COMPARATIVE GENOME ANALYSIS AND EVOLUTIONARY STUDY OF HUMAN PATHOGENIC *YERSINIA* SPECIES

TAN SHI YANG

FACULTY OF DENTISTRY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

COMPARATIVE GENOME ANALYSIS AND EVOLUTIONARY STUDY OF HUMAN PATHOGENIC *YERSINIA* SPECIES

TAN SHI YANG

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF DENTISTRY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2017

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: TAN SHI YANG (I.C/Passport No:

Registration/Matric No: DHA130012

Name of Degree: Doctor of Philosophy

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

COMPARATIVE GENOME ANALYSIS AND EVOLUTIONARY STUDY OF HUMAN PATHOGENIC *YERSINIA* SPECIES

Field of Study: BIOINFORMATICS

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

ABSTRACT

Yersinia is a Gram-negative bacterial genus that includes serious pathogens such as the *Yersinia pestis* which causes plague, and *Yersinia pseudotuberculosis* and *Yersinia enterocolitica* which cause gastrointestinal infections. The remaining species are generally considered to be non-pathogenic to humans. While their virulence mechanisms are well-characterized, the evolution of *Yersinia* pathogens are not well-understood. To understand the evolution of *Yersinia* pathogens and *Yersinia enterocolitica* subspecies, an exhaustive evolutionary and comparative genome studies on a total of 86 *Yersinia* genomes using different bioinformatics approaches were performed. Based on phylogenetic and the gene gain-and-loss analyses, *Yersinia enterocolitica* and *Yersinia pseudotuberculosis*-*Yersinia pestis* were determined as belonging to different phylogroups and have acquired different set of metabolism genes, suggesting that the evolution of human pathogenic *Yersinia* species is most probably triggered by ecological specialization. Besides, pairwise sequence comparisons showed that the *ail* virulence gene of *Yersinia enterocolitica* had higher sequence identities to the *ail* gene family (consists of both *ail* gene and homologs in the same family) of *Yersinia pseudotuberculosis*-*Yersinia pestis* compared to its own *ail* homolog, suggesting that the *ail* gene might have been duplicated in the latter species and then transferred laterally to *Yersinia enterocolitica*. Taken all together, it is proposed that the evolution of *Yersinia* is not in parallel, but rather accompanied by the gene gain-and-loss, gene duplication and lateral gene transfer. This contradicts finding of previous study that suggested the human pathogenic *Yersinia* species might have evolved in parallel to acquire the same virulence determinants.

On the other hand, phylogenetic tree and gene gain-and-loss analyses in this study showed that *Yersinia enterocolitica* strains could be demarcated into three distinct phylogroups,

with each of them acquiring different sets of putative metabolism genes. This postulates that ecological specialization might have triggered subspeciations in *Yersinia enterocolitica* species and lead to the emergence of highly pathogenic, low pathogenic and non-pathogenic subspecies, instead of two subspecies as previously reported. Data gathered in this study also suggest that the lateral gene transfer between subspecies in *Yersinia enterocolitica* might not be extensive as the gene content-based phylogenetic tree highly resembled supermatrix tree. Further virulence gene analyses showed that the *ail* gene was pseudogenized in the non-pathogenic subspecies, probably causing the loss of pYV virulence plasmid and pathogenicity in this subspecies.

To facilitate the ongoing and future research of *Yersinia*, YersiniaBase, a robust and user-friendly *Yersinia* resource and comparative analysis platform for analysing *Yersinia* genomic data was developed. The AJAX-based real-time searching system was implemented to smooth the process of searching genomic data in large databases. YersiniaBase also has in-house developed tools: (1) Pairwise Genome Comparison tool for comparing two user-selected genomes; (2) Pathogenomics Profiling Tool for comparative virulence gene analysis of *Yersinia* genomes; (3) YersiniaTree for constructing phylogenetic tree of *Yersinia*. Successful applications of these useful tools was demonstrated in this study.

Overall, this study provides better insights in elucidating the evolution of human pathogenic *Yersinia* and subspeciation in *Yersinia enterocolitica*. Lastly, the YersiniaBase will offer invaluable *Yersinia* genomic resource and analysis platform for the analysis of *Yersinia* in the future.

ABSTRAK

Yersinia adalah genus bakteria Gram-negatif yang terdiri dari patogen penting seperti *Yersinia pestis* yang menyebabkan wabak, dan *Yersinia pseudotuberculosis* serta *Yersinia enterocolitica* yang menyebabkan jangkitan di usus. Spesies *Yersinia* yang lain tidak patogenik kepada manusia. Walaupun mekanisme kevirulenannya telah difahami, evolusi pathogen *Yersinia* masih kurang difahami. Untuk memahami evolusi patogen *Yersinia* dan subspesies *Yersinia enterocolitica*, kajian menyeluruh ke atas evolusi dan perbandingan genom dalam kalangan 86 genom *Yersinia* telah dijalankan menggunakan pelbagai pendekatan bioinformatik. Berdasarkan analisis filogenetik dan gen gain-and-loss, *Yersinia enterocolitica* dan *Yersinia pseudotuberculosis*-*Yersinia pestis* didapati telah ditempatkan dalam phylogroup yang berlainan dalam pokok filogenetik, dan turut memiliki gen metabolisme yang berbeza. Ini mencadangkan bahawa evolusi patogen *Yersinia* telah dicetuskan oleh pengkhususan ekologi. Di samping itu, perbandingan pasangan jujukan gen menujukkan gen *ail* daripada *Yersinia enterocolitica* mempunyai identiti jujukan gen yang lebih tinggi kepada keluarga gen *ail* daripada *Yersinia pseudotuberculosis*-*Yersinia pestis* berbanding dengan homolog *ail* sendiri. Ini mencadangkan gen *ail* mungkin telah diduplikasi dalam spesies kedua dan dipindahkan ke *Yersinia enterocolitica*. Hasil penemuan ini mencadangkan bahawa evolusi dalam *Yersinia* adalah tidak selari, tetapi dicetuskan oleh gen gain-and-loss, duplikasi gendan pemindahan gen secara lateral. Ini bercanggah dengan hasil kajian sebelum ini yang mana evolusi patogen *Yersinia* dikatakan selari untuk mendapatkan kevirulenan gen yang serupa

Selain itu, analisis pokok filogenetik dan gen gain-and-loss hasil kajian ini menujukkan strain *Yersinia enterocolitica* boleh dibahagikan kepada tiga phylogroup dengan setiapnya memiliki set gen metabolisme yang berbeza. Pengkhususan ekologi telah

dicadangkan sebagai penyebab yang membawa kepada kemunculan subspecies dengan kepatogenan tinggi, kepatogenan rendah dan tidak patogenik, dan bukan dua subspesies seperti yang dilaporkan sebelum ini. Data kajian juga mencadangkan bahawa pemindahan gen secara lateral di antara subspesies *Yersinia enterocolotica* mungkin tidak menyeluruh kerana pokok filogenetik kandungan gennya hampir menyamai pokok filogenetik supermatrix. Analisa lanjut ke atas gen virulen menunjukan gen *ail* telah tersingkir dalam subspesies yang tidak patogenik dan mungkin menyebabkan kehilangan plasmid virulen pYV dan kepatogenan dalam kalangan subspecies ini.

Untuk memudahkan penyelidikan *Yersinia* pada masa kini dan akan datang, YersiniaBase iaitu satu platform untuk sumber dan perbandingan genom *Yersinia* telah dibangunkan Sistem carian real-time berasaskan AJAX ini di implementasikan untuk melancarkan pencarian data dalam pangkalan data yang lebih besar. YersiniaBase dibangunkan dengan alat seperti (1) Pairwise Genome Comparison tool yang membandingkan dua genom *Yersinia* (2) Pathogenomics Profiling Tool yang membandingkan gen virulen *Yersinia* (3) YersiniaTree yang membina pokok filogenetik *Yersinia*. Kejayaan aplikasi system carian ini telah dipamerkan dalam kajian ini.

Kesimpulannya, kajian ini memberikan pemahaman yang lebih baik dalam menjelaskan evolusi *Yersinia* yang patogen kepada manusia dan pembahagian subspesies dalam *Yersinia enterocolitica*. Akhir sekali, YersiniaBase menawarkan sumber genom yang berharga untuk *Yersinia* dan satu platform bagi analisa *Yersinia* pada masa akan datang.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my family members, especially grandparents and parents for their love, spiritual support and motivation of my Ph.D. study. I am grateful to them for giving me strength throughout my study.

I would like to thank to my supervisors, Dr. Choo Siew Woh, Prof. Dr. Irene Tan Kit Ping and Associate Prof. Dr. Fathilah binti Abdul Razak, for their patience, encouragement and insightful comments. Their guidance helped me in all of time of my research and writing of this thesis.

I thank the Chancellery of University of Malaya for providing Bright Sparks Scholarships and High Impact Research Grant (HIR Grant number: UM.C/625/HIR/MOHE/CHAN-08) for supporting my PhD works. Without their precious support, it would not be possible to conduct this study.

To Teo Jing Xian, Athena Ng and Jonathan Tay Weng Chew from other institutes, thank you for always recommending good research articles and having helpful discussions. They provide new insights and enlighten me during my research and thesis writing.

Last but not the least, I would like to thank my friends and colleagues in Genome Informatics Research Laboratory, especially Avirup Dutta and Tan Mui Fern, for their helps.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

AJAX	Asynchronous JavaScript and XML
ANI	Average nucleotide identity
Cas	CRISPR-associated
COG	Cluster of Orthologous Group
CRISPR	Clustered regularly-interspaced short palindromic repeats
CSS	Cascading Style Sheets
HTML	HyperText Markup Language
KOBAS	KEGG Orthology Based Annotation System
LCAY	Last Common Ancestor of All *Yersinia*
LCAHPY	Last Common Ancestor of Human Pathogenic *Yersinia*
NCBI	National Centre for Biotechnology Information
ORF	Open reading frame
PHP	HyperText Preprocessor
RAST	Rapid Annotation using Subsystem Technology
T2SS	Type Two Secretion System
T3SS	Type Three Secretion System
VFDB	Virulence Factors Database
Yop	*Yersinia* outer proteins
Ysc	Yop secretion apparatus

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Overview

Yersinia is a bacterial genus that consists of at least seventeen known species (Clark et al., 2016). Of these species, three species, *Y. enterocolitica*, *Y. pseudotuberculosis* and *Y. pestis* are known human pathogens (Eppinger et al., 2007; Parkhill et al., 2001; Thomson et al., 2006; Wren, 2003). Both *Y. enterocolitica* and *Y. pseudotuberculosis* are foodborne pathogens that cause gastrointestinal disease, whereas the *Y. pestis* is flea-borne pathogen causing catastrophic plague (Eppinger et al., 2007; Parkhill et al., 2001; Thomson et al., 2006; Wren, 2003).

There are many research have identified the key virulence genes underlying the pathogenesis of human pathogenic *Yersinia*, which are harboured in chromosome and pYV virulence plasmid (Cornelis, 2002a; Galindo et al., 2011; Mikula et al., 2012; Yang et al., 1996). Despite of well-studied virulence mechanisms and pathogenesis of *Yersinia*, the evolution of this genus, especially those human pathogenic species, is less focused on. The first model to describe the evolution of human pathogenic *Yersinia* was proposed by Wren (2003). His study suggested that all human pathogenic *Yersinia* shared a common ancestor, which might become pathogenic after the acquisition of pYV plasmid. However, Wren's model did not include human non-pathogenic *Yersinia* species, and was later opposed by Reuter and colleagues, who included both human pathogenic and non-pathogenic *Yersinia* species in their study (Reuter et al., 2014). The authors proposed that ecological specialization caused human pathogenic *Yersinia* to evolve in parallel and acquire the same virulence determinants. Although their hypothesis seems promising, several questions have been raised such as:

- What were the roles or properties of the most recent ancestor shared by the human pathogenic *Yersinia* species?

- How did the hypothetical ancestor cause the ecological specialization?

- Was ecological specialization the only factor that affected the evolution of human pathogenic *Yersinia* species?

In addition to the *Yersinia* genus, narrowing down, there are also disputes in the evolutionary study of *Y. enterocolitica* (Howard et al., 2006). For instance, 16S rRNA sequences were used to classify *Y. enterocolitica* strains into two subspecies: *Y. enterocolitica* subsp. *palearctica* and *Y. enterocolitica* subsp. *enterocolitica* (Neubauer et al., 2000). However, the two-subspecies classification is incongruent with a more recent comparative phylogenomics study of *Y. enterocolitica* proposing the existence of three subspecies in *Y. enterocolitica* (Howard et al., 2006). The proposed subspecies consisted of non-pathogenic lineage, low pathogenic lineage and highly pathogenic lineage. There are also a few questions concerning the subspecies in *Y. enterocolitica* such as:

- What factors have caused the subspeciation in *Y. enterocolitica*?

- Was the most recent ancestor shared by all *Y. enterocolitica* subspecies pathogenic? If yes, what factors have caused the emergence of non-pathogenic lineage? If no, what factors have led to the emergence of pathogenic lineage besides the acquisition of pYV virulence plasmid and several other virulence genes?

- Was the two-subspecies classification accurate because it had previously been reported that 16S rRNA might be unreliable to infer the phylogenetic relationships between *Yersinia* strains (Merhej et al., 2008b)?

As described above, the absence of consensus view has hindered us from fully understanding evolution of human pathogenic *Yersinia* species and subspecies classification of *Y. enterocolitica*. Despite recent larger scale comparative analyses of *Yersinia* species and *Y. enterocolitica* strains (Howard et al., 2006; Reuter et al., 2014), the results/findings are still not comprehensive because there are numerous factors including the ecological specialization, gene gain-and-loss, lateral gene transfer and gene duplication that may play important roles in the evolution of prokaryotes (Jensen, 2001; Lassalle et al., 2015; Ochman et al., 2000; Ravenhall et al., 2015). Therefore, I have performed comparative and evolutionary analyses of *Yersinia* species and the subspecies *Y. enterocolitica* using different bioinformatics approaches in order to explore these factors which are not well-studied.

At the end of this study, I have also developed a specialized comparative analysis platform, designated YersiniaBase, to store the genomic data and provide tools for the comparative analyses of *Yersinia* for research community. YersiniaBase may accelerate the research for those who work on *Yersinia* in future.

1.2 Objectives

The objectives of this study are:

- To perform evolutionary study and comparative analysis on human pathogenic *Yersinia* species and subspecies of *Y. enterocolitica*

- To study the evolutionary factors that caused the emergence of human pathogenic *Yersinia* species and subspecies of *Y. enterocolitica*

- To propose a more complete and robust evolutionary model for human pathogenic *Yersinia* species and subspecies of *Y. enterocolitica*, based on findings from the first two objectives, and compare with current models

- To develop YersiniaBase to store genomic data and provide tools for comparative analyses of *Yersinia*

CHAPTER 2: LITERATURE REVIEW

2.1 The *Yersinia* genus

2.1.1 General properties of *Yersinia*

Yersinia is a Gram-negative bacterium belongs to Enterobacteriaceae family (Williams et al., 2010), consisting of at least seventeen known species such as *Y. pestis*, *Y. pseudotuberculosis*, *Y. enterocolitica*, *Y. aldovae*, *Y. frederiksenii*, *Y. kristensenii*, *Y. ruckeri*, *Y. bercovieri, Y. rohdei*, *Y. intermedia*, *Y. mollaretii*, *Y. massiliensis*, *Y. pekkanenii*, *Y. nurmii*, *Y. aleksiciae*, *Y. wautersii*, and *Y. similis* (Clark et al., 2016). However, there are only three species, *Y. pestis*, *Y. pseudotuberculosis*, *Y. enterocolitica* are known to be pathogenic to humans and one species, *Y. ruckeri* is pathogenic to *Oncorhynchus mykiss* (rainbow trout) (Reuter et al., 2014; Sulakvelidze, 2000; Wren, 2003). *Y. pseudotuberculosis* and *Y. enterocolitica* cause gastrointestinal disease, *Y. pestis* causes plague, whereas *Y. ruckeri* causes enteric redmouth disease in fish (Bottone, 1997; Ewing et al., 1978; Perry & Fetherston, 1997). The rest of the *Yersinia* species are known to be non-pathogenic to living organisms (Reuter et al., 2014; Sulakvelidze, 2000; Wren, 2003).

Overall, taxonomical assignment of each *Yersinia* species is widely accepted except the *Y. ruckeri* that has a controversial taxonomic assignment (Chen et al., 2010; Ewing et al., 1978; Sulakvelidze, 2000). For instance, a previous study showed that *Y. ruckeri* shared similar biochemical activities with *Serratia marcescens* and *Yersinia* species, but it was assigned to *Yersinia* due to the closer guanine-cytosine content (Ross et al., 1966).

2.1.2 Virulence genes of human pathogenic *Yersinia*

Pathogenesis is due to the presence of virulence genes in bacteria, which are responsible for causing disease in the host (Chen et al., 2012). One of the key factors in the pathogenesis of human pathogenic *Yersinia* is the deployment of the pYV virulence plasmid (Cornelis, 2002a) in the human pathogenic *Y. pestis*, *Y. pseudotuberculosis* and *Y. enterocolitica*. The Type Three Secretion System (T3SS) encoded by the pYV plasmid is transcribed into two components: *Yersinia* outer proteins (Yop) and Yop secretion apparatus (Ysc) (Cornelis, 2002a, 2002b). When the direct contact between pathogenic *Yersinia* and mammalian cell is established, the pathogen uses Ysc to inject Yop effectors into host cell. The Yop effectors are able to take over the signalling system of the host cell, paralyze the host cell, and allow the bacteria to escape phagocytosis (Cornelis, 2002a; Felek et al., 2010; McDonald et al., 2003; Navarro et al., 2005).

Besides the *ysc-yop* T3SS locus, the chromosome-borne *inv* (invasin), *ail* (attachment-invasion locus), *psa* (pH 6 antigen) locus, and pYV plasmid-borne *yad*A (*Yersinia* adhesion) are also important virulence genes to human pathogenic *Yersinia* (Cornelis et al., 1998; Felek et al., 2010; Grassl et al., 2003; Iriarte & Cornelis, 1995; Mikula et al., 2012). These genes allow *Yersinia* to adhere and invade into the host cell, induce agglutination, resist to human serum, and assist in the Yop delivery (Cornelis et al., 1998; Felek et al., 2010; Grassl et al., 2003; Iriarte & Cornelis, 1995; Mikula et al., 2012).

While the abovementioned virulence genes are found in every human pathogenic *Yersinia*, high pathogenicity island that harbours *ybt* (abbreviation of yersiniabactin) locus encoding yersiniabactin synthesis, transport and uptake system, is found only in highly pathogenic *Yersinia* species (Carniel, 2001; Heesemann, 1987). Yersiniabactin is a type of siderophore which enables highly pathogenic *Yersinia* to scavenge iron in iron-limited

environment (Carniel, 2001; Carniel et al., 1996). The importance of yersiniabactin, of which it is able to compete iron with host cell, had been shown in experiments using mice as models whereby the presence of yersiniabactin can increase the virulence of *Yersinia* species and cause the death of mice (de Almeida et al., 1993; Heesemann, 1987; Pelludat et al., 2002).

2.1.3 *Yersinia pseudotuberculosis* and *Yersinia pestis*

Both *Y. pseudotuberculosis* and *Y. pestis* are known human pathogens. A previous evolutionary study has identified *Y. pestis* is a very recent descendant from *Y. pseudotuberculosis*, as recent as 1,500–20,000 years ago (Achtman et al., 1999). Despite of their close evolutionary relationships, they deploy a totally different infection routes and exhibit distinct virulence traits (Chain et al., 2004). For instance, the *Y. pseudotuberculosis* is a food-borne human enteropathogen that causes Far East scarlet-like fever and yersiniosis (Eppinger et al., 2007), whereas the *Y. pestis* is a flea-transmitted systematic pathogen that causes plague (Green et al., 2014; Perry & Fetherston, 1997).

Previous studies had also revealed many changes taking place in the genome of *Y. pestis* since its divergence from the *Y. pseudotuberculosis* (Chain et al., 2004; Hinnebusch et al., 2002; Parkhill et al., 2001). Besides acquiring pPla (or pPst) and pFra plasmids that are not found in the enteropathogenic *Y. pseudotuberculosis*, the *Y. pestis* also has frameshift mutations in the *inv* and *yad*A virulence genes (Parkhill et al., 2001). These genomic alternations are thought to be important to its lifestyle, which is the change from food-borne transmission (in *Y. pseudotuberculosis*) to flea-borne transmission (in *Y. pestis*) (Hinnebusch et al., 2002). The pFra plasmid encodes phospholipase D, allowing the *Y. pestis* to survive inside flea's gut (Hinnebusch et al., 2002). When the flea that carries *Y.*

pestis bites on next infected host, the pathogen will enter the host body through subcutaneous site. At this point, the pPla plasmid which encodes plasminogen activator allows *Y. pestis* to disseminate from the initial infection site (Caulfield & Lathem, 2012; Lathem et al., 2007).

On the other hand, the *inv* and *yad*A genes are lost in *Y. pestis* and these genes are thought to be not required for flea-borne infection route (Simonet et al., 1996; Skurnik & Wolf-Watz, 1989). Both *inv* and *yad*A are functional and playing important role in *Y. enterocolitica* and *Y. pseudotuberculosis* as they are still needed for oral-route infection. (Mikula et al., 2012).

2.1.4 *Yersinia enterocolitica*

The *Y. enterocolitica* is a foodborne enteropathogen that causes yersiniosis (Bottone, 1997; Galindo et al., 2011). Despite exhibiting similar virulence traits with *Y. pseudotuberculosis*, the *Y. enterocolitica* is not genetically similar with *Y. pseudotuberculosis* and they are evolutionarily distant to each other (Thomson et al., 2006).

Biochemical tests have categorized *Y. enterocolitica* strains into six biogroups, namely biogroup-1A, biogroup-1B, biogroup-2, biogroup-3, biogroup-4 and biogroup-5 (Bottone, 1997; Wauters et al., 1987). These biogroups can be further categorized by their geographical location as well as pathogenicity level (Batzilla et al., 2011b; Bottone, 1997; Thomson et al., 2006). For instance, biogroup-1A does not have pYV virulence plasmid and is generally considered to be non-pathogenic; biogroup-1B has pYV plasmid as well as high pathogenicity island which harbours *ybt* locus and is highly pathogenic; biogroup-2, biogroup-3, biogroup-4 and biogroup-5 have pYV plasmid but do not possess high pathogenicity island and is low pathogenic (Bottone, 1997; Carniel et al., 1996; Pelludat

et al., 1998). On top of that, biogroup-1B is prevalent in North America, while the rest is prevalent in Europe (Batzilla et al., 2011b; Thomson et al., 2006).

At the taxonomical level, *Y. enterocolitica* strains have been classified into two subspecies, namely *Y. enterocolitica* subsp. *enterocolitica* and *Y. enterocolitica* subsp. *palearctica* based on their 16S rRNA sequences (Neubauer et al., 2000). The subspecies classification corresponds to their geographic distribution, whereby *Y. enterocolitica* subsp. *enterocolitica* is mainly found in North America while *Y. enterocolitica* subsp. *palearctica* is prevalent in Europe (Neubauer et al., 2000). In 2006, phylogenomics study of *Y. enterocolitica* had been performed using DNA microarrays and comparative genomic approaches (Howard et al., 2006). However, these two scientific works are incongruent to each other. For instance, Howard and colleagues found that there were three subspecies existed within *Y. enterocolitica*, corresponding to highly pathogenic, low pathogenic and non-pathogenic biotypes (Howard et al., 2006). Besides that, they hypothesized that: (1) highly pathogenic lineage was a direct descendant of the last common ancestor of *Y. enterocolitica* (2) separation of highly pathogenic lineage with the other two lineages, which were low pathogenic and non-pathogenic, might be due to biogeographic movement (Howard et al., 2006).

2.1.5 Evolution of human pathogenic *Yersinia*

In addition to independent studies of *Y. pseudotuberculosis*-*Y. pestis* and *Y. enterocolitica*, there are also evolutionary studies consisting of all three human pathogenic *Yersinia* species, albeit the number of these studies is lesser (Reuter et al., 2014; Wren, 2003). One of the earliest models to elucidate the evolution of human pathogenic *Yersinia* has been documented by Wren (Wren, 2003). Wren proposed that the *Y. enterocolitica*, *Y. pseudotuberculosis* and *Y. pestis* shared a common pathogenic ancestor, which had

acquired pYV plasmid. However, Wren's study did not include the other non-pathogenic *Yersinia* species and has contradicted to another model proposed by Reuter and co-workers (Reuter et al., 2014). The latter study hypothesized that early ecological specialization has separated human pathogenic *Yersinia* into different lineages, causing the human pathogenic *Yersinia* to evolve in parallel, but acquired the similar virulence genes.

2.2 Evolutionary study in prokaryotes

2.2.1 Phylogenetic studies

The small subunit ribosomal RNA, which is also known as 16S rRNA, has been the standard to generate phylogenetic tree and perform taxonomic classification of prokaryotes due to its presence in every bacterial genome (Rajendhran & Gunasekaran, 2011; Woese et al., 1990). However, discrepancies between 16S rRNA phylogenetic tree and phylogenetic trees derived from other genes, such as 23S rRNA and housekeeping genes, had been reported in *Helicobacter* and *Yersinia* (Dewhirst et al., 2005; Merhej et al., 2008a). In *Yersinia*, phylogenetic tree inferred by using housekeeping genes, which included *rpo*B, *hsp*60, *gyr*B and *sod*A, was found to be more congruent with biochemical test result compared to 16S rRNA (Merhej et al., 2008a).

Besides the 16S rRNA gene, core genes (i.e., genes that present in all genomes) has also been proposed to be an alternative approach to infer phylogenetic relationships between bacteria (Daubin et al., 2002). The use of core genes is on the basis that lateral gene transfer seldom affects bacterial core genes (Daubin et al., 2002). In several studies, supermatrix tree, which is based on concatenation of a set of core genes, can produce phylogenetic tree with good accuracy (de Queiroz & Gatesy, 2007; Lapierre et al., 2014; Tonini et al., 2015; von Haeseler, 2012).

Besides using sequences, the gene content (i.e., presence and absence of gene families in a given list of genomes) can also be used for bacterial phylogenetic tree construction (Snel et al., 1999). In this method, the evolutionary distance between genomes is calculated based on their shared gene content or number of shared genes: higher number of shared genes leads to shorter evolutionary distance and vice versa. The pioneer of this approach also argued that lateral gene transfer does not have extensive impact on the gene content of bacterial genomes (Snel et al., 1999). Indeed, another study showed that vertical inheritance, rather than lateral gene transfer, formed the dominant process in bacterial evolution and determined its gene content (Snel et al., 2002).

2.2.2 Ecological specialization

Prokaryotes can evolve and diversify by adapting to different ecological niches (Cohan, 2002). This process is termed ecological specialization or ecological speciation, whereby distinguishable lineages or populations can arise due to the acclimatization in distinct niches, and independent evolution between each other (Kopac et al., 2014). Hence, the gain of new genes to adapt to new niches, and the loss of ancestral genes to live in more restricted niches, appear to be important in ecological speciation (Lassalle et al., 2015). Despite of this, recombination can still be ongoing at early phase of ecological speciation, often at loci that do not bring advantage to the niche survival (Cadillo-Quiroz et al., 2012). If the recombination is extensive, it forms cohesive forces between populations and constrains divergence of lineages. Nevertheless, recombination in most bacteria might not be as frequent as usually thought (Cohan, 2001). As mutations still play a dominant role in shaping bacterial genomes, the rate of recombination tends to decrease as nucleotide divergence forms the barrier to the process (Fraser et al., 2007; Majewski et al., 2000).

The ecological specialization in *Yersinia* was recently proposed by Reuter and colleagues (Reuter et al., 2014). For instance, they found *Y. enterocolitica* is specialized in utilizing cobalamin, 1,2-propanediol, tetrathionate and hydrogen due to the presence of metabolism genes in its genome to exploit these compounds. On the other hand, the pathogenic counterparts, *Y. pseudotuberculosis* and *Y. pestis* do not have these metabolism genes. They concluded in the study that the early adaptation to different ecological niches have split human pathogenic *Yersinia* into several lineages.

2.2.3 Gene gain-and-loss

The gene content in bacterial genome follows the phyletic pattern due to differential gene gain and gene loss between the lineages of a given phylogenetic tree (Snel et al., 1999). Hence, the gene gain-and-loss analysis can be used to predict ancestral gene content, acquired and lost genes along lineages (Csuros, 2010). Due to its ability to reconstruct ancestral events, the approach had been widely used in evolutionary study to infer how bacterial lineages diversified and evolved in the past as well as to predict the factors which led to the emergence of pathogens (Desai et al., 2013; Georgiades et al., 2011; Kettler et al., 2007).

In *Yersinia*, the most popular example to describe gene gain-and-loss is the emergence of *Y. pestis* from *Y. pseudotuberculosis* (Achtman et al., 1999; Chain et al., 2004). As described above, *Y. pestis* has acquired two additional plasmids which are not found in its ancestor and transformed into a more catastrophic human pathogen. Another example which applied gene gain-and-loss analysis would be the evolutionary study of *Prochlorococcus* (Kettler et al., 2007). In the study, the phylogenetic tree constructed by Kettler and colleagues could cluster *Prochlorococcus* strains into high-light adapted and low-light adapted clade, which also corresponded to different ecotypes. Through the

study of gained and lost genes, they found several genes which exclusive to two different clades could define their distinct traits. For instance, several genes that present only in high-light adapted clade could be up regulated when the intensity of light is high.

2.2.4 Lateral gene transfer

Lateral gene transfer is the non-vertical exchange of DNA between bacterial cells via conjugation, transformation or transduction (Ochman et al., 2000). Genes which can be transferred using such mechanisms include antibiotic resistance genes, virulence genes and metabolic genes (Ochman et al., 2000; Pal et al., 2005). Example of laterally transferred virulence locus is T3SS locus harboured by *Salmonella typhimurium*, which could be transferred by bacteriophage (Mirold et al., 1999). The O-antigen gene cluster of *Y. kristensenii* O11 was also acquired in lateral, from either *Escherichia*, *Salmonella*, or *Klebsiella* (Cunneen & Reeves, 2007).

Besides increasing the diversity in the bacterial genomes, the laterally transferred genes enable the recipient genomes to adapt to new ecological niches (Marri et al., 2007; Wiedenbeck & Cohan, 2011). These adaptive genes consist of both single gene and genomic islands, which up to hundreds of kilo-base pairs (Marri et al., 2007). For instance, *Escherichia coli*, a Gram-negative bacterium, has acquired *gap*C (glyceraldehyde-3-phosphate dehydrogenases) from Gram-positive bacteria, allowing them to adapt to aquatic environment (Espinosa-Urgel & Kolter, 1998).

Several approaches could be used to infer lateral gene transfer events including the construction of phylogenetic trees to look for discrepancy, the calculation of guanine-cytosine content across genome sequences, and searching for organisms located within the top BLAST hits (Ravenhall et al., 2015). For instance, a previous study has

successfully used the top BLAST hits approach to discover extensive lateral gene transfer between *Thermotoga maritima* (bacteria) and Archaea (Nelson et al., 1999).

2.2.5 Orthologs and paralogs

Orthologs and paralogs are two different terms assigned to genes which duplicate in different time. When a gene is originated from the same ancestor and duplicates during speciation, it is called ortholog, otherwise it is called paralog (Jensen, 2001). As orthologs involve in the divergence of lineages and present in each genome, it can be used to infer the evolutionary relationships between lineages. A common approach derived from this concept is to use single copy core gene (i.e., gene which is present in only one copy in all genomes) or concatenation of these genes to construct phylogenetic tree or supermatrix tree (Daubin et al., 2002; de Queiroz & Gatesy, 2007; Segata & Huttenhower, 2011). This approach is made on the basis of orthologs and core genes have evolutionarily similar history. In a case where an ortholog duplicates after the speciation, the duplicated genes, i.e. the paralogs, will be co-orthologous to the ortholog in the counterpart lineage which also diverged from the same ancestor through speciation (Sonnhammer & Koonin, 2002).

Unlike the orthologs or core genes, paralogs are not related to speciation and it could be only present in some genomes, but missing in the rest. Thus, the paralogs are not suitable to infer phylogenetic relationships between lineages (Jensen, 2001; Sonnhammer & Koonin, 2002). Despite of this limitation, paralogs can give evolutionary advantages to the bacteria. When there are two duplicated genes present in the bacterial genome, the cell has redundant copies of the same gene which encodes for the same function (Wagner, 2002). This scenario results in lower pressure of purifying selection in one of the paralogs (Kondrashov et al., 2002; Wagner, 2002). Mutations are then allowed to be accumulated in one of them, while another gene can still perform the same physiological role as before

(Wagner, 2002), leading to the rise of beneficial mutations and novel functions (Kondrashov et al., 2002; Wagner, 2002). For instance, a recent study has shown that there were gene duplications of Leucine rich repeat gene family which contributed to the evolution of human pathogenic *Leptospira* (Xu et al., 2016).

2.2.6 Clustered Regularly-interspaced Short Palindromic Repeats

The Clustered Regularly-interspaced Short Palindromic Repeats (CRISPR) is a locus found in bacterial genome. It is an array consists of repetitive DNA repeats and spacer sequences (Horvath & Barrangou, 2010). The function of CRISPR is to protect bacteria against foreign DNA materials such as prophage and plasmid sequence (Horvath & Barrangou, 2010; Makarova et al., 2011; Nozawa et al., 2011). The genes that are responsible to this immunity are located adjacent to CRISPR, designated as *cas* (CRISPR-associated) (Horvath & Barrangou, 2010). In general, bacteria may acquire immunity against a specific phage or plasmid by capturing and integrating fragments of foreign DNA inside CRISPR array. The novel sequence is known as spacer. The spacer will provide resistance when bacteria encounter the same sequence again in the future. The immunity process is carried out by matching spacer to the foreign DNA sequence using Cas proteins (Makarova et al., 2011). Previous experiments have shown that the CRISPR-Cas can interfere lateral gene transfer by restricting the transfer of antibiotic resistance genes among pathogens as well as the conjugative plasmids in bacteria (Marraffini, 2013; Marraffini & Sontheimer, 2008).

2.3 Microbial genome databases

In recent years, a new trend of collecting bacterial genomes into a single database has emerged as an effective way to analyse their genomes. Consequently, many specialized genomic databases have been developed, especially for human disease pathogens. There are a number of databases, such as "Microbial Genome Database for Comparative Analysis", "Integrated Microbial Genomes" and "Pathosystems Resource Integration Center", which provide a wide array of microbial genomes for comparative genomics (Markowitz et al., 2012; Uchiyama et al., 2013; Wattam et al., 2014). However, they do not provide functionalities for comparing and visualizing the virulence gene profiles of user-selected *Yersinia* strains. These databases also do not provide the option for comparative virulence gene analysis based on the virulence genes of the strains. Moreover, most of these existing platforms also lack of user-friendly web interfaces which allows real-time and fast querying and browsing of genomic data.

CHAPTER 3: METHODOLOGY

3.1 Genome sequences retrieval and annotation

A total of 197 genome sequences were downloaded from the National Centre for Biotechnology Information (NCBI) public database (Benson et al., 2015). 86 of them were *Yersinia*, two were *Haemophilus influenza*, and the rest were other genus within Enterobacteriaceae. The accession number and details of *Yersinia* genomes are tabulated in Appendix A and Table 3.1 respectively.

Table 3.1: List of *Yersinia* genomes used in this study with their corresponding isolation source and geographical area. Human pathogenic strains are coloured in red.

Species name	Strain name	Isolation source	Geographic area
Y. aldovae	670-83	Fish	Norway
Y. aleksiciae	159	Human faeces	Finland
Y. frederiksenii	Y225	Unknown	Unknown
Y. intermedia	Y228	Unknown	Unknown
Y. kristensenii	Y231	Unknown	Unknown
Y. rohdei	YRA	Animal faeces	Germany
Y. ruckeri	YRB	Fish liver	Unknown
Y. ruckeri	Big Creek 74	*Oncorhynchus tshawytscha*	Oregon, United States
Y. similis	228	Rabbit	Germany
Y. enterocolitica	ERL073947	Sheep	New Zealand
Y. enterocolitica	IP2222	Unknown	Unknown
Y. enterocolitica	NFO	Unknown	Unknown
Y. enterocolitica	ERL08708	Human	New Zealand
Y. enterocolitica	YE13/03	Human faeces	United Kingdom
Y. enterocolitica	IP26014	Bovine	France
Y. enterocolitica	YE53/30444	Pig	Germany
Y. enterocolitica	SZ662/97	Human	Germany
Y. enterocolitica	IP26618	Chicken meat	Italy
Y. enterocolitica	YE208/02	Pig	United Kingdom
Y. enterocolitica	ERL053435	Human	New Zealand
Y. enterocolitica	H1527/93	Human	Germany
Y. enterocolitica	YE53/03	Human case	United Kingdom
Y. enterocolitica	YE30/03	Human case	United Kingdom
Y. enterocolitica	YE41/03	Human case	United Kingdom
Y. enterocolitica	YE15/07	Human case	Germany
Y. enterocolitica	ERL053484	Avian	New Zealand
Y. enterocolitica	YE35/02	Human case	United Kingdom
Y. enterocolitica	YE69/03	Human case	United Kingdom

Table 3.1: List of *Yersinia* genomes used in this study with their corresponding isolation source and geographical area. Human pathogenic strains are coloured in red, continued.

Species name	Strain name	Isolation source	Geographic area
Y. enterocolitica	YE77/03	Pig	United Kingdom
Y. enterocolitica	IP27818	Human stool	France
Y. enterocolitica	YE46/02	Cattle	United Kingdom
Y. enterocolitica	YE228/02	Pig	United Kingdom
Y. enterocolitica	YE38/03	Human case	United Kingdom
Y. enterocolitica	YE04/02	Sheep	United Kingdom
Y. enterocolitica	YE205/02	Pig	United Kingdom
Y. enterocolitica	YE09/03	Human faeces	United Kingdom
Y. enterocolitica	YE13/02	Cattle	United Kingdom
Y. enterocolitica	YE221/02	Pig	United Kingdom
Y. enterocolitica	YE227/02	Pig	United Kingdom
Y. enterocolitica	ATCC 9610	Homo sapiens	New York
Y. enterocolitica	E701	Human stool	Unknown
Y. enterocolitica	8081	Human blood	United States
Y. enterocolitica	ST5081	Unknown	Unknown
Y. enterocolitica	SC9312-78	Human	United States
Y. enterocolitica	E736	Human stool	Unknown
Y. enterocolitica	WA	Human blood	United States
Y. enterocolitica	WA-314	Human blood	Unknown
Y. enterocolitica	Y286	Unknown	United States
Y. enterocolitica	SZ5108/01	Human	Germany
Y. enterocolitica	SZ375/04	Human	Germany
Y. enterocolitica	SZ506/04	Human	Germany
Y. enterocolitica	IP05342	Hare	Belgium
Y. enterocolitica	IP00178	Hare	United Kingdom
Y. enterocolitica	IP26042	Cattle	France
Y. enterocolitica	IP06077	Hare	France
Y. enterocolitica	YE3094/96	Animal	Europe
Y. enterocolitica	YE04/03	Human faeces	United Kingdom
Y. enterocolitica	YE238/02	Pig	United Kingdom
Y. enterocolitica	IP20322	Milk	Greece
Y. enterocolitica	YE153/02	Cattle	United Kingdom
Y. enterocolitica	IP26249	Human stool	France
Y. enterocolitica	YE149/02	Sheep	United Kingdom
Y. enterocolitica	YE213/02	Pig	United Kingdom
Y. enterocolitica	Y11	Human	Germany
Y. enterocolitica	IP26656	Human stool	France
Y. enterocolitica	PhRBD_Ye1	Swine	Philippines
Y. enterocolitica	YE12/03	Human stool	United Kingdom
Y. enterocolitica	YE07/03	Human faeces	United Kingdom
Y. enterocolitica	IP 10393	Homo sapiens	France
Y. enterocolitica	105.5R(r)	Human	China
Y. enterocolitica	Y127	Unknown	Unknown
Y. enterocolitica	YE74/03	Human case	United Kingdom
Y. enterocolitica	YE237/02	Pig	United Kingdom
Y. enterocolitica	YE214/02	Pig	United Kingdom

Table 3.1: List of *Yersinia* genomes used in this study with their corresponding isolation source and geographical area. Human pathogenic strains are coloured in red, continued.

Species name	Strain name	Isolation source	Geographic area
Y. enterocolitica	YE212/02	Pig	United Kingdom
Y. enterocolitica	YE218/02	Pig	United Kingdom
Y. enterocolitica	YE56/03	Human case	United Kingdom
Y. enterocolitica	IP21447	Pig stool	England
Y. enterocolitica	YE119/02	Sheep	United Kingdom
Y. enterocolitica	1127	Human	Ireland
Y. enterocolitica	2/C/53NMD7	Pig	Ireland
Y. enterocolitica	YE11/03	Human case	United Kingdom
Y. pseudotuberculosis	IP31758	Human patient	Primorski, Soviet Union
Y. pseudotuberculosis	IP32953	Human patient	France
Y. pestis	KIM10+	Human	Kurdistan, Iran
Y. pestis	CO92	Human	United States

All genomes were annotated by using Rapid Annotation using Subsystem Technology (RAST) online server to obtain list of open reading frames (ORFs), coding sequences and protein sequences (Aziz et al., 2008). Function of each protein sequence was predicted by using BLASTP and HMM to search against four databases, including Cluster of Orthologous Group (COG), Virulence Factors Database (VFDB), KEGG Orthology Based Annotation System (KOBAS) and TIGRFAMs (Altschul et al., 1990; Chen et al., 2012; Galperin et al., 2015; Haft et al., 2013; Johnson et al., 2010; Xie et al., 2011).

3.2 Calculation of average nucleotide identity

JSpecies was used to calculate average nucleotide identity (ANI) in *Yersinia* chromosomes and pYV plasmids (Richter & Rossello-Mora, 2009). The pairwise ANI values were manually inspected to find the highly similar groups of *Yersinia* genomes.

3.3 Protein sequence clustering

The 197 downloaded genome sequences were categorized into three datasets which are described in Table 3.2. ProteinOrtho was used to cluster protein sequences of each dataset independently using default parameters: 1E-5 as E-value cut-off, 25% as minimum percentage of identity and 50% as minimum percentage of sequence coverage (Lechner et al., 2011).

Table 3.2: Categorization of 197 genome sequences into three datasets together with their respective outgroup.

Dataset name	Genomes	Outgroup
Enterobacteriaceae	▪ *Y. aldovae* 670-83 ▪ *Y. aleksiciae* 159 ▪ *Y. enterocolitica* Y11 ▪ *Y. enterocolitica* 8081 ▪ *Y. frederiksenii* Y225 ▪ *Y. intermedia* Y228 ▪ *Y. kristensenii* Y231 ▪ *Y. pestis* CO92 ▪ *Y. pestis* KIM10+ ▪ *Y. pseudotuberculosis* IP31758 ▪ *Y. pseudotuberculosis* IP32953 ▪ *Y. rohdei* YRA ▪ *Y. ruckeri* YRB ▪ *Y. ruckeri* Big Creek 74 ▪ *Y. similis* 228 ▪ Other genus in Enterobacteriaceae	▪ *H. influenzae* 86-028NP ▪ *H. influenzae* Rd KW20
Yersinia	▪ *Y. aldovae* 670-83 ▪ *Y. aleksiciae* 159 ▪ *Y. enterocolitica* Y11 ▪ *Y. enterocolitica* 8081 ▪ *Y. frederiksenii* Y225 ▪ *Y. intermedia* Y228 ▪ *Y. kristensenii* Y231 ▪ *Y. pestis* CO92 ▪ *Y. pestis* KIM10+ ▪ *Y. pseudotuberculosis* IP31758 ▪ *Y. pseudotuberculosis* IP32953 ▪ *Y. rohdei* YRA ▪ *Y. ruckeri* YRB ▪ *Y. ruckeri* Big Creek 74 ▪ *Y. similis* 228	▪ *S. liquefaciens* HUMV-21 ▪ *S. liquefaciens* ATCC 27592
Y. enterocolitica	All 73 genome sequences of *Y. enterocolitica*	*Y. kristensenii* Y231

3.4 Multiple sequence alignment

Protein sequences of each single copy core gene family from all three datasets were aligned using L-INS-i algorithm, which was implemented in Multiple Alignment using Fast Fourier Transform (MAFFT) program (Katoh & Standley, 2013). The aligned protein sequences were then translated back to codon sequences using PAL2NAL (Suyama et al., 2006). Poorly aligned regions of codon sequences were removed using GBlocks (Castresana, 2000).

3.5 Estimation of recombination

PHI program was used to estimate the probability of recombination in each aligned codon sequence, with 10,000 iterations and 0.05 as p-value cut-off (Bruen et al., 2006). Non-recombinant codon sequences from each single copy core gene family were concatenated to form a "non-recombinant super-sequence" (de Queiroz & Gatesy, 2007).

Without recombination estimation by PHI, the aligned codon sequences from each single copy core gene family were also concatenated to form "super-sequence". ClonalFrameML was used to estimate the rate of recombination to mutation in the super-sequence (Didelot & Wilson, 2015).

3.6 Phylogenetic tree and network construction

For *Y. enterocolitica* and Enterobacteriaceae datasets, FastTree2 was used to construct supermatrix tree based on their respective aligned non-recombinant super-sequence (Price et al., 2010). While for *Yersinia* dataset, RAxML was used to construct supermatrix tree based on non-recombinant super-sequence (Stamatakis, 2014). Both FastTree2 and RAxML were set to use GTR+GAMMA model, maximum likelihood method with 1,000 bootstrap iterations.

MEGA6 was used to reconstruct another neighbour-joining phylogenetic tree based on gene content (i.e., the presence and absence of gene in each family) as described previously (Snel et al., 1999; Tamura et al., 2013).

SplitsTree was used to reconstruct the phylogenetic network based on super-sequence (without recombination testing) from *Y. enterocolitica* dataset (Huson & Bryant, 2006).

3.7 Gene gain-and-loss analysis

Count (a bioinformatics tool) was used to reconstruct gene gain and gene loss events in Enterobacteriaceae and *Y. enterocolitica* datasets with maximum parsimony. Acquired or lost genes in ancestors of interest were inspected manually (Csuros, 2010).

3.8 Clustered Regularly-interspaced Short Palindromic Repeats analysis

CRISPR Recognition Tool was used to predict CRISPR array, which consists of spacers and repetitive sequences, in each *Yersinia* genome (Bland et al., 2007). BLASTN was then used to search spacer against NCBI database to predict the sources of the spacer sequences (Clark et al., 2016).

3.9 *inv* homolog analysis

Nucleotide and protein sequences of known *inv* in *Y. enterocolitica* 8081 were retrieved from VFDB and were used as reference in this study (Chen et al., 2012). BLASTP was used to search for genes homologous to reference *inv* in each *Yersinia* genome (Altschul et al., 1990). The BLASTP output was filtered based on 1E-7 as E-value and 50% sequence completeness of query and subject sequences. Query start and stop positions from each BLASTP alignment were manually inspected. ProteinOrtho was used to cluster protein sequences of *inv* homologs into gene family using default parameters: 1E-5 as E-

value cut-off, 25% as minimum percentage of identity and 50% as minimum percentage of sequence coverage (Lechner et al., 2011). Protein sequences of *inv* homologs in each gene family were aligned using L-INS-i implemented in MAFFT (Katoh & Standley, 2013). Aligned proteins sequences were translated back into codon sequences using PAL2NAL (Suyama et al., 2006). PHI was then used to estimate probability of recombination in each aligned codon sequence (Bruen et al., 2006).

3.10 *ail* homolog analysis

Nucleotide and protein sequences of known *ail* in *Y. pestis* CO92 were retrieved from VFDB and were used as reference in this study (Chen et al., 2012). TBLASTN and BLASTP were used to search for genes homologous to reference *ail* in each *Yersinia* genome (Altschul et al., 1990). The BLASTP output was filtered based on 1E-7 as E-value and 50% sequence completeness of query and subject sequences. ProteinOrtho was used to cluster protein sequences of *ail* homologs into gene family using default parameters: 1E-5 as E-value cut-off, 25% as minimum percentage of identity and 50% as minimum percentage of sequence coverage (Lechner et al., 2011).

Putative pseudogenized *ail* was predicted based on previous study (Lerat & Ochman, 2004). Briefly, nucleotide region mapped by TBLASTN was extracted and aligned with reference *ail* coding sequence using G-INS-i implemented in MAFFT (Katoh & Standley, 2013). ExPASy web server was used to translate the extracted region into protein sequence (Gasteiger et al., 2003). Both aligned coding sequence and protein sequence were compared to identify putative frameshift mutation and premature stop codon.

3.11 Development of YersiniaBase

All available *Yersinia* genomes, which included both draft and complete genomes, were downloaded from the NCBI database and annotated using RAST (Aziz et al., 2008; Benson et al., 2015). PSORTb version 3.0 was used to determine the subcellular localization of the proteins predicted by the RAST (Yu et al., 2010). Hydrophobicity and molecular weight of the RAST-predicted proteins were computed using in-house developed Perl scripts.

A relational database was implemented using MySQL version 14.12. All of the biological data of *Yersinia* strains were rearranged to fit into the designed database schema and stored in the MySQL database. A web interface was built using HyperText Markup Language (HTML), HyperText Preprocessor (PHP), JavaScript, Cascading Style Sheets (CSS) and "asynchronous JavaScript and XML" (AJAX). CodeIgniter version 2.1.3, a popular PHP framework was used in order to a provide model-view-controller, which can separate application data, presentation and background logic and process into three distinct modules, allowing the source codes and *Yersinia* biological data to be arranged in a clear and organized manner, indirectly allowing easier future updating of YersiniaBase.

Several in-house designed bioinformatics tools were developed and integrated into YersiniaBase. Python, Perl and R languages were used to develop PGC for comparing between two genomes through global alignment, PathoProT for generating heat map to visualize presence and absence of virulence genes in selected genomes and YersiniaTree to generate phylogenetic tree of *Yersinia* strains based on their housekeeping genes and 16S rRNA. These three popular scripting languages create complex pipelines that perform back-end calculations of the bioinformatics tools, aided communications between the web server and the application server and also easier transfer of data between the web server and the application server.

CHAPTER 4: RESULTS (PART 1): THE HUMAN PATHOGENIC *YERSINIA* SPECIES

4.1 Properties of *Yersinia* genomes

Both *Yersinia* and Enterobacteriaceae datasets were used to study *Yersinia* genus and human pathogenic *Yersinia* species in this section. Summary of the annotation of fifteen *Yersinia* complete genomes used in this section are tabulated in Table 4.1.

Table 4.1: Summary of genome annotation of *Yersinia* species used in this study. Human pathogenic *Yersinia* strains are coloured in red.

Species	Strain	Genome size (base pair)	Guanine-cytosine content (%)	Total coding sequence	Total rRNA operon	Total tRNA
Y. aldovae	670-83	4,471,090	47.69	3,985	7	82
Y. aleksiciae	159	4,000,307	49.04	3,569	7	75
Y. enterocolitica	Y11	4,553,420	47.01	4,155	7	70
Y. enterocolitica	8081	4,615,899	47.27	4,167	7	81
Y. frederiksenii	Y225	4,495,532	47.40	4,016	7	82
Y. intermedia	Y228	4,859,749	47.47	4,320	7	81
Y. kristensenii	Y231	4,496,569	47.40	4,012	7	81
Y. pestis	KIM10+	4,600,755	47.64	4,033	7	73
Y. pestis	CO92	4,653,728	47.64	4,090	6	70
Y. pseudotuberculosis	IP31758	4,723,306	47.54	4,013	7	86
Y. pseudotuberculosis	IP32953	4,743,972	47.61	4,072	7	85
Y. rohdei	YRA	4,372,253	47.03	3,791	7	81
Y. ruckeri	YRB	3,605,216	47.50	3,162	7	79
Y. ruckeri	Big Creek 74	3,699,725	47.64	3,268	7	81
Y. similis	228	4,903,722	46.97	4,327	7	87

The genome size of *Yersinia* species ranged from the smallest 3,605,216 base pairs in *Y. ruckeri* to the largest 4,903,722 base pairs in *Y. similis*. Despite there was difference of more than one million base pairs between *Yersinia* genomes, all *Yersinia* species were highly similar in guanine-cytosine percentage and number of rRNA operon. *Y. ruckeri* had the least number of coding sequences (CDSs) or genes, whereas the *Y. similis* had the most number of genes.

4.2 Average nucleotide identity between *Yersinia* genomes

To study the genomic similarity between *Yersinia* species, average nucleotide identity (ANI) was calculated between their chromosomes and between their pYV virulence plasmids, separately. The pairwise ANI values between each pair of *Yersinia* chromosomes are tabulated in Table 4.2.

Table 4.2: ANI values (in percentage) between each pair of *Yersinia* chromosomes. Pairwise ANI values between human pathogenic *Yersinia* strains are highlighted in red.

	Y. aldovae 670-83	*Y. aleksiciae* 159	*Y. enterocolitica* 8081	*Y. enterocolitica* Y11	*Y. frederiksenii* Y225	*Y. intermedia* Y228	*Y. kristensenii* Y231	*Y. pestis* CO92	*Y. pestis* KIM10+	*Y. pseudotuberculosis* IP31758	*Y. pseudotuberculosis* IP32953	*Y. rohdei* YRA	*Y. ruckeri* Big Creek 74	*Y. ruckeri* YRB	*Y. similis* 228
Y. aldovae 670-83	---	82.96	84.09	83.93	83.09	83.19	83.24	81.45	81.49	81.50	81.40	82.11	77.67	77.66	81.35
Y. aleksiciae 159	82.98	---	83.71	83.47	83.78	84.06	83.75	81.77	81.85	81.78	81.80	83.25	78.06	77.99	81.79
Y. enterocolitica 8081	84.02	83.79	---	96.62	87.49	83.54	87.51	81.84	81.78	81.60	81.61	83.99	77.70	77.65	81.55
Y. enterocolitica Y11	83.89	83.46	96.72	---	87.50	83.23	87.49	81.22	81.29	81.26	81.37	83.73	77.66	77.61	81.25
Y. frederiksenii Y225	83.07	83.79	87.43	87.45	---	83.45	100.00	81.48	81.39	81.45	81.32	83.95	77.50	77.54	81.39
Y. intermedia Y228	83.34	84.26	83.60	83.34	83.57	---	83.51	81.80	81.74	81.63	81.62	82.91	77.85	77.85	81.68
Y. kristensenii Y231	83.07	83.79	87.44	87.44	100.00	83.46	---	81.47	81.39	81.44	81.32	83.96	77.51	77.54	81.37
Y. pestis CO92	81.25	81.86	81.65	81.20	81.21	81.47	81.34	---	99.94	98.78	99.13	80.98	77.89	77.81	94.53
Y. pestis KIM10+	81.24	81.85	81.64	81.20	81.20	81.46	81.34	99.92	---	98.79	99.14	80.99	77.91	77.81	94.56
Y. pseudotuberculosis IP31758	81.34	81.83	81.49	81.17	81.13	81.45	81.39	98.93	98.94	---	99.04	80.91	77.90	77.78	94.70
Y. pseudotuberculosis IP32953	81.25	81.84	81.53	81.30	81.22	81.45	81.37	99.18	99.19	98.97	---	80.93	77.95	77.75	94.61
Y. rohdei YRA	82.09	83.26	83.90	83.70	83.89	82.67	83.90	80.94	81.01	80.84	80.92	---	77.76	77.87	80.99
Y. ruckeri Big Creek 74	77.56	78.01	77.48	77.41	77.48	77.68	77.56	77.81	77.86	77.80	77.74	77.79	---	97.68	77.66
Y. ruckeri YRB	77.66	77.95	77.60	77.45	77.50	77.67	77.57	78.05	78.09	77.76	77.71	77.84	97.63	---	77.81
Y. similis 228	81.28	81.81	81.45	81.17	81.13	81.42	81.37	94.59	94.60	94.74	94.63	80.90	77.78	77.77	---

I found that the pairwise ANI values between *Y. pestis* and *Y. pseudotuberculosis* chromosomes were very high, with at least 98.78% identity, probably because the *Y. pestis* is a very recent descendant of *Y. pseudotuberculosis* (Achtman et al., 1999). To describe the close relationship between the two species, I will refer both *Y. pestis* and *Y. pseudotuberculosis* as *Y. pseudotuberculosis*-*Y. pestis* throughout this thesis. Despite *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis* are the only human pathogenic species within *Yersinia* genus, I found that the ANI between them were only approximately 81%. My data also clearly showed that the *Y. enterocolitica* was more similar to human nonpathogenic *Y. frederiksenii* and *Y. kristensenii* (approximately 87% identity), whereas the *Y. pseudotuberculosis*-*Y. pestis* were closer to the nonpathogenic *Y. similis* (approximately 94% identity). These findings suggest that the pathogenic *Y. enterocolitica* is not closely related to the *Y. pseudotuberculosis*-*Y. pestis*, which is consistent with the findings reported in previous studies (McNally et al., 2016; Thomson et al., 2006).

Besides the ANI of *Yersinia* chromosomes, the ANI between the pYV virulence plasmids harboured by human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis* were also calculated (Table 4.3).

Table 4.3: ANI values (in percentage) between the pYV virulence plasmids harboured by human pathogenic *Yersinia* species.

	Y. pestis CO92	*Y. pseudotuberculosis* IP32953	*Y. enterocolitica* 8081	*Y. enterocolitica* Y11
Y. pestis CO92	---	99.33	97.58	96.96
Y. pseudotuberculosis IP32953	99.49	---	97.21	96.62
Y. enterocolitica 8081	97.65	97.54	---	98.20
Y. enterocolitica Y11	97.50	97.38	98.20	---

Unlike the low chromosomal ANI values (approximately 81%) between the *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*, I found that their pYV plasmids were highly similar with approximately 97% identity, suggesting that their pYV plasmids are closely related to each other, and might have the same origin albeit they are borne by distantly related human pathogenic *Yersinia* species.

4.3 *Yersinia* gene families

4.3.1 Gene families in *Yersinia* chromosomes

To study the gene families in *Yersinia*, all chromosomal protein sequences of the fifteen *Yersinia* complete genomes were clustered into 8,943 gene families. Percentage of orthologous, co-orthologous, dispensable and strain-specific gene families are shown in Figure 4.1.

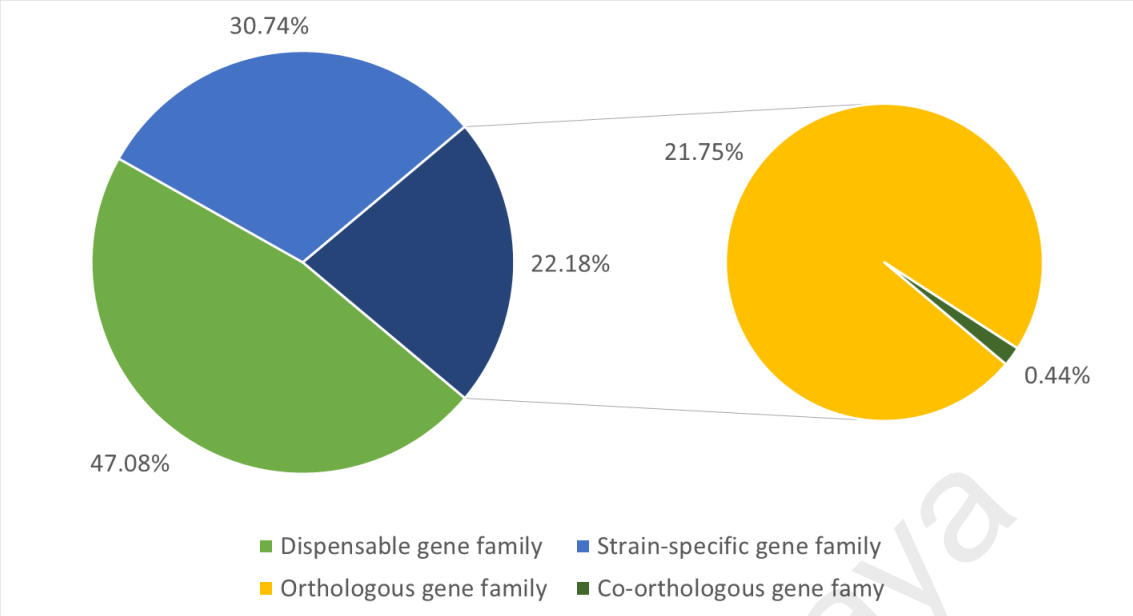

Figure 4.1: Percentage of orthologous, co-orthologous, dispensable and strain-specific gene families present in the *Yersinia* genomes.

Given a gene family, ortholog is the gene which duplicated during speciation and is present in exactly one copy in each genome whereas co-ortholog is also an ortholog, but it has been duplicated in some of the genomes (Lechner et al., 2011). I found that the orthologous and co-orthologous gene families contributed less than a quarter (22.18%) to the total gene families of *Yersinia*. This suggests that most of the *Yersinia* species had acquired new gene families in the past, diluting the proportion of (co-)orthologous genes. Gene duplication might not be the dominant process in *Yersinia* as I found that less than 1% of total gene family were co-orthologous gene family.

On the other hand, I also found that dispensable and strain-specific gene families had contributed more than three quarters (77.82%) to the total gene family of *Yersinia*. Given a gene family, dispensable gene is the gene which present in at least two genomes but not all whereas strain-specific gene, as the name implies, the gene can be only found in one strain (Tettelin et al., 2005). The large number of dispensable and strain-specific genes in *Yersinia* suggests that the genomes of *Yersinia* are likely mosaic and frequently change to acclimatize to new environments (Segerman, 2012). For instance, certain lineages

might acquire new genes, which are not present in other lineages, to survive in new niches, or they might lose some genes which no longer required. In either case, the process could lead to dispensable gene families, whereby the genes will be only found in certain lineages.

4.3.2 Gene families in the pYV virulence plasmids

To study the gene families in pYV virulence plasmids harboured by human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*, the protein sequences encoded by their pYV plasmids were clustered into 128 gene families. Percentage of core, dispensable and strain-specific gene families from the pYV plasmids are shown in Figure 4.2.

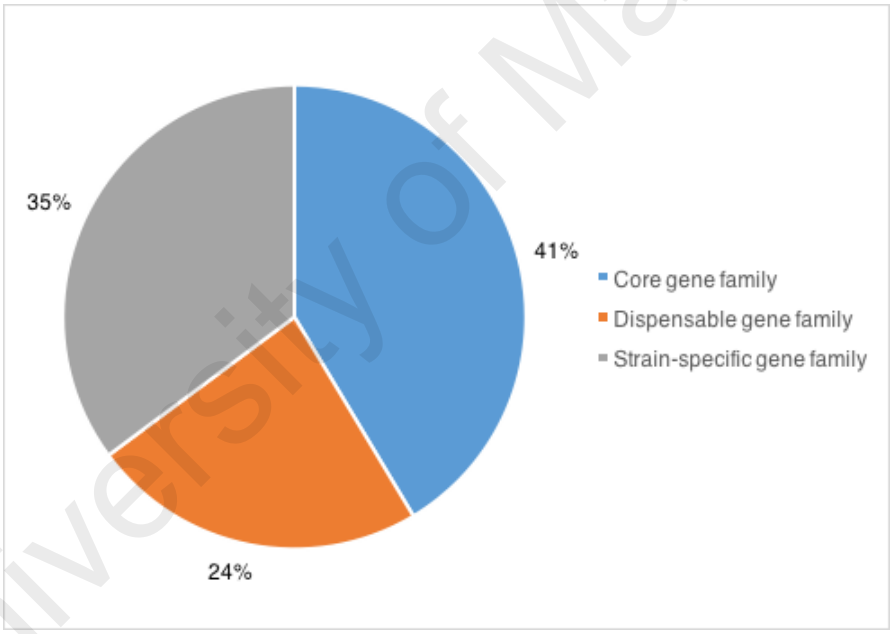

Figure 4.2: Percentage of orthologous, co-orthologous, dispensable and strain-specific gene families present in pYV virulence plasmids harboured by human pathogenic *Yersinia* species.

Despite only four pYV plasmids were used in this analysis, I found that the core gene (gene which is present in every plasmid) contributed less than half (41%) to the gene pool of pYV plasmids, suggesting that they are not conserved across all human pathogenic *Yersinia* species. Similarly to chromosomes, the pYV might have experienced multiple gene gain or gene loss events in *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis* throughout the evolution time. Nevertheless, I found that the core genes mainly encoded

for Ysc-Yop T3SS, which is the most important virulence factors deployed by human pathogenic *Yersinia* species to infect host cells (Cornelis, 2002a).

On the other hand, transposase and integrase genes were found in the dispensable gene families. This further suggests that insertion of other mobile genetic elements into pYV plasmids had been taken place in the past, which decreased the proportion of pYV core genes. Among the strain-specific genes, I found *ars*BCRH (arsenic detoxification genes) in *Y. enterocolitica* Y11. This is consistent with a previous study suggesting that the presence of *ars* locus might be important for the spread of low pathogenic *Y. enterocolitica* Y11 (Neyt et al., 1997).

4.4 Phylogenetic relationships between *Yersinia* and *Yersinia ruckeri*

The taxonomic classification of fish pathogenic *Y. ruckeri* is still uncertain since its discovery (Chen et al., 2010; Sulakvelidze, 2000). Construction of Enterobacteriaceae supermatrix tree might be helpful in resolving this uncertainty. By inferring from 141,057 nucleotides, I have constructed an Enterobacteriaceae supermatrix tree, which is shown in Figure 4.3. The supermatrix tree clearly showed that *Y. ruckeri* was clustered together with the rest *Yersinia* species, but not *Serratia*.

To provide support to the visual inspection, I have also calculated the branch length, which was based on number of nucleotide substitutions per site, between *Y. ruckeri* and the other species. *Y. ruckeri* YRB was used as the reference and the results are tabulated in Table 4.4. From the calculation, *Yersinia* was the nearest species to *Y. ruckeri*, followed by *Hafnia alvei* and *Serratia*. This confirms that *Y. ruckeri* is belonged to *Yersinia* genus.

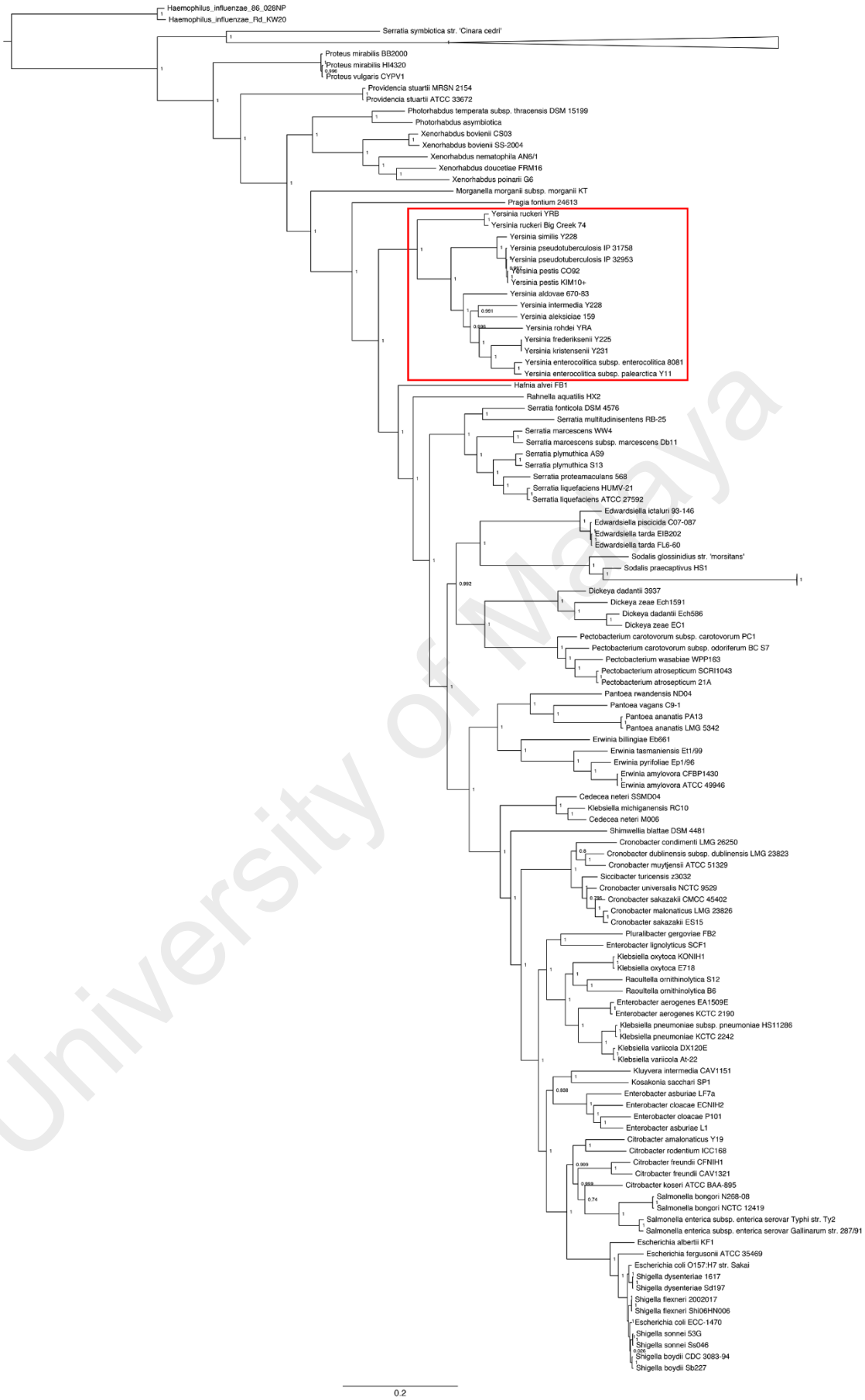

Figure 4.3: Enterobacteriaceae supermatrix tree constructed using non-recombinant super-sequence with 141,057 nucleotides and rooted by *Haemophilus influenzae*. *Yersinia* genus was bordered by red.

Table 4.4: First 30 species nearest to *Y. ruckeri* YRB (reference) based on calculation of branch length.

Genome	Branch length
Yersinia ruckeri YRB	0
Yersinia ruckeri Big Creek 74	0.01359
Yersinia similis Y228	0.28015
Yersinia pseudotuberculosis IP 31758	0.28028
Yersinia aldovae 670-83	0.28067
Yersinia pseudotuberculosis IP 32953	0.28068
Yersinia pestis CO92	0.2823
Yersinia pestis KIM10	0.2823
Yersinia intermedia Y228	0.29822
Yersinia aleksiciae 159	0.29836
Yersinia enterocolitica Y11	0.30546
Yersinia frederiksenii Y225	0.30556
Yersinia kristensenii Y231	0.30556
Yersinia enterocolitica 8081	0.30667
Yersinia rohdei YRA	0.30795
Hafnia alvei FB1	0.42284
Serratia plymuthica S13	0.44202
Serratia marcescens WW4	0.44206
Serratia plymuthica AS9	0.44284
Serratia marcescens subsp. *marcescens* Db11	0.44351
Rahnella aquatilis HX2	0.44453
Serratia fonticola DSM 4576	0.44644
Serratia liquefaciens ATCC 27592	0.45524
Serratia liquefaciens HUMV-21	0.45577
Serratia proteamaculans 568	0.45618
Serratia multitudinisentens RB-25	0.49778
Pragia fontium 24613	0.5042
Cedecea neteri SSMD04	0.53695
Pectobacterium carotovorum subsp. *carotovorum* PC1	0.53755
Klebsiella michiganensis RC10	0.55149

4.5 Phylogenetic relationships between *Yersinia* species

In this section, I studied the phylogenetic relationships of *Yersinia* species using different bioinformatics approaches such as supermatrix tree and gene content-based phylogenetic tree.

4.5.1 *Yersinia* supermatrix tree

A non-recombinant super-sequence with length of 245,662 nucleotides was used to infer *Yersinia* supermatrix tree and to study the phylogenetic relationships between 15 *Yersinia* genomes (Figure 4.4).

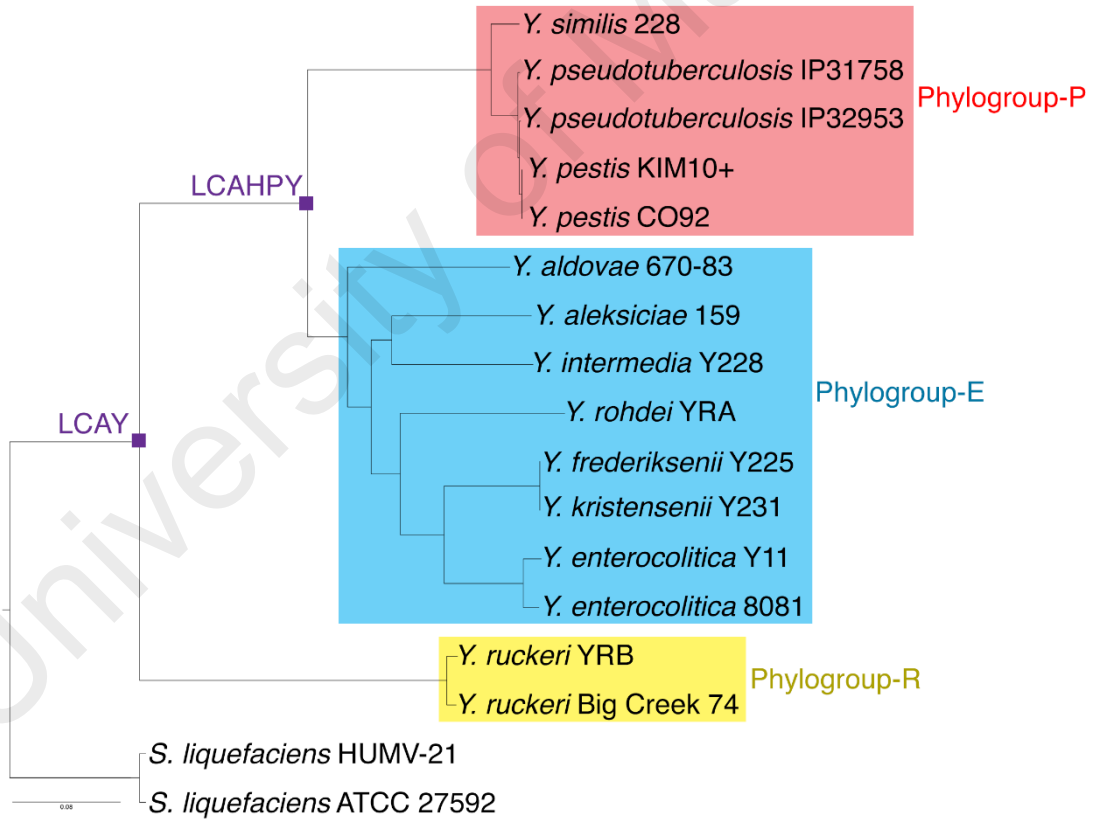

Figure 4.4: *Yersinia* supermatrix tree inferred from non-recombinant super-sequence and rooted by *Serratia liquefaciens*. All *Yersinia* species descended from the "Last Common Ancestor of all *Yersinia*" (LCAY) while human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis* shared the "Last Common Ancestor of Human Pathogenic *Yersinia*" (LCAHPY). Phylogroup-P, phylogroup-E and phylogroup-R were highlighted by magenta, cyan and yellow respectively. All internal nodes had bootstrap value of 100.

The supermatrix tree demarcated fifteen *Yersinia* genomes into three phylogroups, namely phylogroup-P, phylogroup-E and phylogroup-R. Generally, all *Yersinia* species diverged from "Last Common Ancestor of all *Yersinia*" (LCAY). The fish pathogenic *Y. ruckeri* was isolated from the rest species and belonged to phylogroup-R. My data clearly showed that the human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis* did not belong to the same phylogroup. *Y. enterocolitica* belonged to phylogroup-E, and it was grouped together with the human non-pathogenic *Y. aldovae*, *Y. aleksiciae*, *Y. intermedia*, *Y. rohdei*, *Y. frederiksenii* and *Y. kristensenii*. On the other hand, *Y. pseudotuberculosis*-*Y. pestis* belonged to phylogroup-P and grouped together with *Y. similis*, another human non-pathogenic *Yersinia* species. As *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis* appeared at the basal position of the supermatrix tree and closer to the non-pathogenic species in their respectively phylogroup, they might have evolved from different non-pathogenic populations.

I also found that the human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis* did not share the same direct ancestor, but instead, their common ancestor, namely "Last Common Ancestor of Human Pathogenic *Yersinia*" (LCAHPY), was far from them and closer to LCAY. In general, the *Yersinia* supermatrix tree suggests that *Y. enterocolitica* is distantly related to *Y. pseudotuberculosis*-*Y. pestis* and their time of divergence could be very ancient.

4.5.2 *Yersinia* gene content-based phylogenetic tree

Based on the information of the presence and absence of gene families in each genome, a gene content phylogenetic tree was reconstructed to infer the phylogenetic relationship of *Yersinia* species in Figure 4.5.

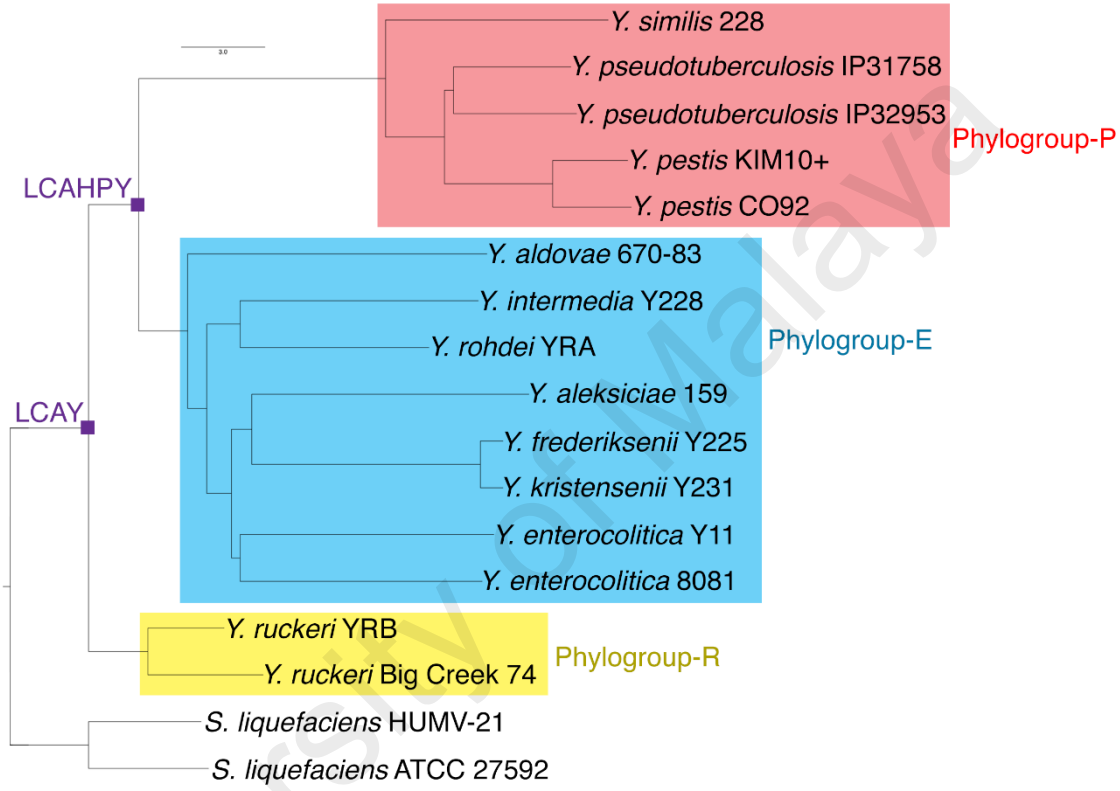

Figure 4.5: *Yersinia* gene content-based phylogenetic tree reconstructed based on the information of the presence and absence of gene families in each genome. The tree exhibits highly similar phyletic patterns with supermatrix tree whereby the genomes were grouped into phylogroup-R, phylogroup-E and phylogroup-P.

I found that the gene content-based phylogenetic tree had highly similar phyletic patterns with the *Yersinia* supermatrix tree, whereby all fifteen genomes were also grouped into three phylogroups. Besides that, LCAY and LCAHPY, which were present in the supermatrix tree could also be recovered in the gene content-based phylogenetic tree. These observations suggest that there are distinguishable gene contents harboured by *Yersinia* species from different phylogroups, probably due to phylogroup-specific gene turnovers.

4.6 Recombination in *Yersinia*

To examine whether recombination is extensive within *Yersinia*, I have estimated the rate of recombination compared to the mutation rate using super-sequence present in all *Yersinia* species. The average relative rate of recombination (R) to mutation (θ) of *Yersinia* genus was estimated to be $R/\theta = 0.011$, mean DNA import length was $\delta = 603$ base pairs, mean divergence of imported DNA was $\nu = 0.041$. As R/θ was smaller than 1, mutation is likely a dominant occurrence in the genus, taking place at 90 (1/0.011=90) times more often than recombination. It is possible that the recombination across different species would decrease due to the increase of nucleotide divergence between *Yersinia* species (Majewski et al., 2000).

4.7 Gene gain-and-loss in *Yersinia*

To understand how gene content was changed since the emergence of *Yersinia*, gene gain-and-loss analysis was performed. Acquired and lost genes in the ancestors were predicted based on maximum parsimony algorithm applying on gene content of present-day *Yersinia* species. The number of gene gain, gene loss and number of gene in each ancestor are shown in Figure 4.6.

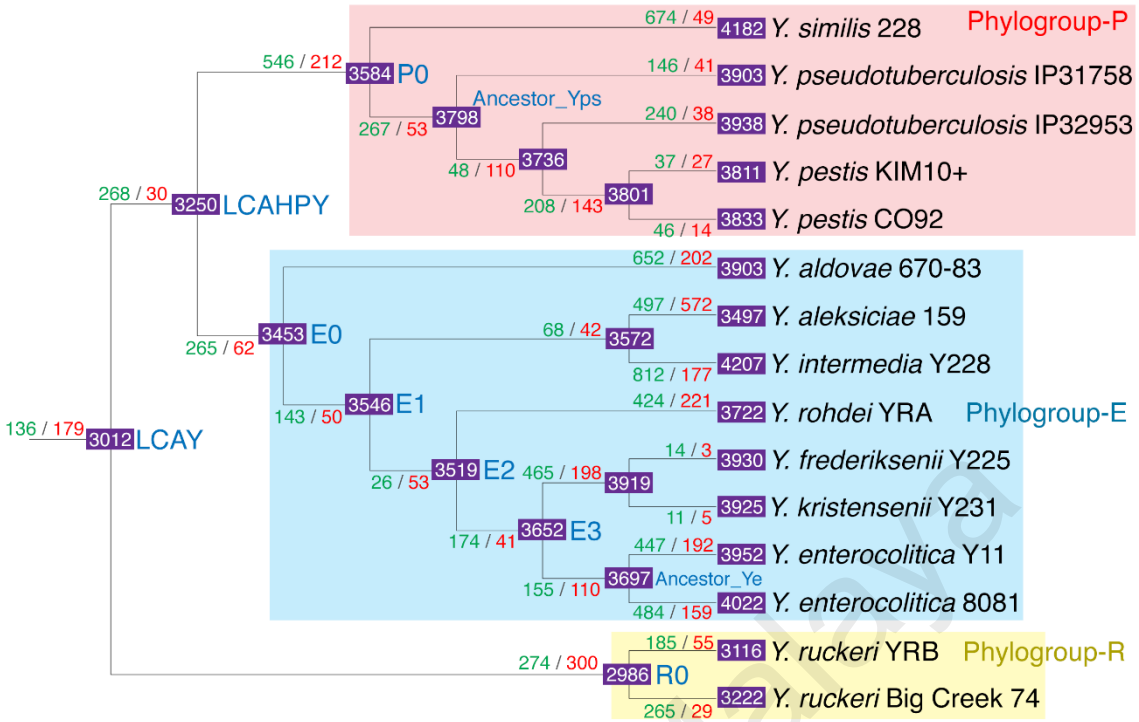

Figure 4.6: *Yersinia* cladogram showing the reconstruction of gene gain-and-loss in ancestral nodes. Green, red, white colour numbers indicate gene gain, gene loss and number of gene in each ancestor respectively. Hypothetical ancestors of interest are labelled in blue colour text.

My analysis showed that the gene gain was dominant in the evolution of *Yersinia*. For instance, I found that the number of gene gain was greater than the number of gene loss in most ancestor and present-day genomes. This suggests that the genomes of *Yersinia* generally increase in size. In the subsections below, I shall discuss the acquired and lost genes in ancestors of interest which leading to the emergence of human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*. To ease the discussion that followed, I have assigned hypothetical names for ancestors of interest in the Figure 4.6.

4.7.1 Emergence of Last Common Ancestor of all *Yersinia* (LCAY)

LCAY was hypothesized as the most recent hypothetical ancestor shared by all *Yersinia* species in this study. LCAY might have preferred an aerobic environment due to the acquisition of aerobic citrate transporter genes (*tct*ABCDE) (Brocker et al., 2009). It might be able to extract heme from the host organism as indicated by the gain of heme receptor gene (*has*R) and hemophore gene (*has*A) (Letoffe et al., 1999).

I found that the LCAY had lost *dsd*ACX which are D-serine tolerance genes. D-serine is an anti-microbial compound abundant in the brain and urinary tract and it is able to inhibit growth of enterohemorrhagic *Escherichia coli* (Connolly et al., 2016). I hypothesize that the LCAY might be unable to survive in brain and urinary tract due to the loss of these important genes.

4.7.2 Emergence of last common ancestor of fish pathogenic *Yersinia ruckeri* (R0 ancestor)

R0 was hypothesized as the most recent hypothetical ancestor shared by all fish pathogenic *Y. ruckeri* strains in phylogroup-R. It was the direct descendant of LCAY. I found that R0 ancestor had acquired several virulence genes, which including *ysa*-T3SS, *yts*1-Type Two Secretion System (T2SS) and *ent*ABCES. Previous studies have shown that both *ysa* and *yts*1 loci are found only in high pathogenic *Y. enterocolitica* while *ent* locus, which encodes for ruckerbactin (a type of siderophore), is up regulated when *Y. ruckeri* infects fish (Fernandez et al., 2004; Haller et al., 2000; Iwobi et al., 2003). The acquisitions of these virulence genes suggest they are important to the pathogenesis of *Y. ruckeri* strains in rainbow trout. R0 ancestor had also gained *rsb*W (anti-anti-sigma factor) and *rsb*V (anti-anti-sigma factor), which play an important role in osmoprotection of

Streptomyces coelicolor, probably reflecting the importance of these genes to *Y. ruckeri* since it lives in freshwater (Lee et al., 2004).

I found several metabolism and transporters genes were lost in R0 ancestor. For instance, *iol*ABCDEG (myo-inositol degradation genes) that encode enzymes to degrade myo-inositol, an abundant compound in soil, was lost in R0. The loss of *iol* locus probably due to a narrower niche in *Y. ruckeri* as it mainly associates with and infects fishes (Sulakvelidze, 2000). Besides, the loss of *efe*UOB (acid-induced ferrous transporter genes) suggests the shift of *Y. ruckeri* to freshwater that has more neutral pH resulting the genes unnecessary for survival (Cao et al., 2007).

4.7.3 Emergence of Last Common Ancestor of all human pathogenic *Yersinia* species (LCAHPY)

LCAHPY was the most recent ancestor shared by human pathogenic *Y. pseudotuberculosis*-*Y. pestis* and *Y. enterocolitica*. Therefore, gene gain-and-loss in LCAHPY would be interesting to study. I found that LCAHPY had acquired *pga*ABCD (poly-beta-1,6-N-acetyl-D-glucosamine synthesis and transport genes), *pel* and *pel*W (pectate lyases), *tog*BANM and *tog*T (oligogalacturonide transporter genes). Previous studies have shown that these genes allow human enteric pathogen, such as *Escherichia coli* EDL933, to persist and proliferate on vegetables (Hugouvieux-Cotte-Pattat & Reverchon, 2001; Roy et al., 1999; Yamazaki et al., 2011; Yaron & Romling, 2014). Hence, the acquisition of *pga*, *pel* and *tog* loci suggests that LCAHPY might have acquired the capability to grow on vegetables and be introduced into the human gastrointestinal tract after consumption of vegetables.

Besides the genes for surviving the outside human intestines, I found that the LCAHPY ancestor had also acquired genes such as *yut* and *urt*ABCDE (urea transporter genes),

*nix*A and *ynt*ABCDE (nickel transporter genes), *ure*ABCEFGD (urease genes). Previous study showed that these genes allow *Helicobacter pylori* to colonize and cause infections in stomach (Mobley, 1996). This suggests that those loci might allow colonization of LCAHPY in gastrointestinal tract after consumption of contaminated food by the host. The survival of LCAHPY in human gastrointestinal tract could be further enhanced through the acquisition of *lsr*ABCD (autoinducer-2 transporter genes) and *lsr*ABCD (autoinducer-2 processing enzymes genes). For instance, previous study proposed that enteric bacteria might use Lsr proteins to interrupt intercellular communication among competing bacterial cells (Xavier et al., 2007).

4.7.4 Emergence of Phylogroup-E

LCAHPY, which shared by both human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*, has generally diverged into two lineages: phylogroup-E and phylogroup-P. Subsequent speciation events in phylogroup-E eventually gave to the rise of enteropathogenic *Y. enterocolitica* (Figure 4.4 and Figure 4.6). Many hypothetical ancestors had emerged during the speciation events before *Y. enterocolitica*, namely E0, E1, E2 and E3 in this study. I found that these hypothetical ancestors had acquired *hyb* and *hyf* loci (hydrogenase genes), *cbi*ABCDETFGHJKLMNOQP and *cob*STU (cobalamin biosynthesis genes), *pdu*BCELPQW (1,2-propanediol degradation genes) and *ttr*ABCRS locus (tetrathionate reduction genes). Previous study showed that these loci can provide growth advantage to *Salmonella enterica* serotype Typhimurium in the gastrointestinal tract and to outcompete other enteric bacteria (Rohmer et al., 2011). This may suggest similar role of these acquired genes during the emergence of phylogroup-E species. Besides that, I suggest that cellobiose might be important to the metabolism of phylogroup-E species as they gained second copy of *cel*ABC (cellobiose phosphotransferase system).

4.7.5 Emergence of human pathogenic *Yersinia enterocolitica* in phylogroup-E

The abovementioned E3 ancestor further diverged into two lineages; one of them was "the last common ancestor of all *Y. enterocolitica* strains" (Ancestor_Ye). Cellobiose seemed to be important to the lifestyle of Ancestor_Ye because I found that it had acquired the third copy of *cel*ABC genes. *rut*RABCDEFG (pyrimidine catabolism genes) were also acquired by Ancestor_Ye but their physiological role in bacteria is not yet understood. Nevertheless, the absence of *rut* locus in the all non-pathogenic species within phylogroup-E suggests that it might play an important role in the virulence traits of *Y. enterocolitica*. Most importantly, I found that the Ancestor_Ye had acquired pYV virulence plasmids and *myf* genes (Iriarte & Cornelis, 1995; Miller et al., 1989).

4.7.6 Emergence of Phylogroup-P

Phylogroup-P was the sister lineage of phylogroup-E. Both of them diverged from the LCAHPY. Subsequent speciation events in phylogroup-P gave to the rise of human pathogenic *Y. pseudotuberculosis*. From the supermatrix tree (Figure 4.4 and Figure 4.6), P0 ancestor, which descended directly from LCAHPY, had emerged before "the last common ancestor of all *Y. pseudotuberculosis*-*Y. pestis* strains" (Ancestor_Yps). I found that P0 had gained *ter*ZABCD (tellurite resistance genes) and *rip*ABC (itaconate catabolism genes). Itaconate is a type of antimicrobial compound secreted by macrophages. Previous studies have shown that these two loci are adaptive strategies for bacteria to survive inside macrophages, suggesting similar role for phylogroup-P species (Ponnusamy & Clinkenbeard, 2015; Ponnusamy et al., 2011; Sasikaran et al., 2014). Besides adaptive genes, P0 ancestor had also gained several virulence genes, including *pil*WVUSRQPONML (type IV pilus gene cluster which resides in *Yersinia* adhesion pathogenicity island), *psa*ABCEF (pH 6 antigen genes). All of these virulence genes have

been shown to be important in pathogenicity of human pathogenic *Yersinia* (Collyn et al., 2002; Yang et al., 1996).

On the other hand, *bcs*GFE and *bcs*QABZC, which are cellulose synthesis genes, were lost in P0 ancestor. A recent study has demonstrated that repression of cellulose biosynthesis in *Salmonella* when it is inside a macrophage could increase its virulence (Pontes et al., 2015). It is possible that the loss of cellulose biosynthesis genes and gain of itaconate catabolism genes could enhance survival of phylogroup-P species inside the macrophage.

4.7.7 Emergence of human pathogenic *Yersinia pseudotuberculosis* in phylogroup-P

All *Y. pseudotuberculosis* strains descended from "the last common ancestor of all *Y. pseudotuberculosis*-*Y. pestis* strains" (Ancestor_Yps). I found that Ancestor_Yps ancestor had acquired *mqs*R and *mqs*A, which are a pair of toxin-antitoxin genes. Previous study has showed that *mqs*R and *mqs*A are the most highly up regulated gene in persistent *E. coli* cells and they regulated other physiological genes (Brown et al., 2009). This suggests that the *mqs* toxin-antitoxin gene pair may be important for the pathogenic phylogroup-P species to overcome stresses from the host immune mechanisms.

4.8 *inv* homologs in *Yersinia*

Invasin, which encoded by the *inv* gene, is one of the virulence factors deployed by human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis* to attach and invade host cell lining (Mikula et al., 2012). However, the evolutionary history of *inv* is less focused compared to its virulence mechanism. Hence, I attempted to search for *inv* homologs in *Yersinia* to understand its evolution. Using the functional *inv* from *Y. enterocolitica* 8081 as reference, I found that there were a total of 13 (including reference) *inv* homologs in *Yersinia* species and clustered them into a single gene family based on at least 25% sequence identity, 50% sequence completeness and E-value < 1E-5 (Table 4.5.).

Table 4.5: BLASTP output where the functional *inv* of *Y. enterocolitica* 8081 was used as reference query to search for homologs in *Yersinia*. The functional *inv* genes of human pathogenic species are highlighted in red.

Subject	Subject locus tag	Subject start	Subject end	Query start	Query end	Sequence identity (%)
Y. aldovae 670-83	CP009781.1_CDS_159	58	811	46	808	40.67
Y. aldovae 670-83	CP009781.1_CDS_407	159	619	69	540	46.33
Y. enterocolitica 8081	AM286415.1_CDS_2507	1	835	1	835	100
Y. enterocolitica Y11	FR729477.2_CDS_1378	1	835	1	835	99.04
Y. frederiksenii Y225	CP009364.1_CDS_1102	131	653	69	623	36.56
Y. kristensenii Y231	CP009997.1_CDS_515	131	653	69	623	36.56
Y. pestis CO92	AL590842.1_CDS_1781	7	467	245	730	46.26
Y. pestis KIM10+	AE009952.1_CDS_2457	7	467	245	730	46.26
Y. pseudotuberculosis IP31758	CP000720.1_CDS_2233	1	746	1	730	51.28
Y. pseudotuberculosis IP32953	BX936398.1_CDS_1693	1	746	1	730	51.41
Y. rohdei YRA	CP009787.1_CDS_1058	162	736	55	640	39.07
Y. rohdei YRA	CP009787.1_CDS_1742	139	624	54	544	49.3
Y. similis Y228	CP007230.1_CDS_3220	1	751	1	730	51.28

Gene gain and loss analysis suggests that the *inv* homologs might have been inherited from LCAHPY. By analysing the BLASTP outputs, I found that the *inv* homologs were generally present in most *Yersinia* species. However, it should be noted that the *inv* gene has previously been reported to be pseudogenized in human pathogenic *Y. pestis* (Simonet et al., 1996). On the other hand, two human non-pathogenic species, *Y. aldovae* and *Y. rohdei,* might have duplicated the *inv* homolog. Recombination testing showed that the probability of recombination is $p < 0.01$, proposing that the recombination might have been taken place in the past. In general, *inv* homologs which were inherited from the LCAHPY might have undergone different evolutionary changes such as gene loss, gene duplication and recombination.

Besides, I found some differences in the query and subject start positions in the alignments when comparing between the human pathogenic *Yersinia* and non-pathogenic species: the aligned regions between *inv* homologs of all non-pathogenic *Yersinia* (except *Y. similis*) and reference *inv* did not start at first amino acid. This might account for different expression of the protein transcribed from *inv* homolog in non-pathogenic *Yersinia* species as the N-terminal of Inv is important for proper localization in the outer membrane of *Yersinia* as previously reported (Leong et al., 1990).

4.9 *ail* homologs in *Yersinia*

Ail is another virulence factor deployed by human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis* besides Inv (Mikula et al., 2012). Similarly with the *inv* gene, the evolutionary history of the *ail* gene is also less focused compared to its virulence mechanism. Using the functional *ail* from *Y. pestis* as reference, I found that there were a total of 32 *ail* homologs present in *Yersinia* and clustered them into three gene families based on at least 25% sequence identity, 50% sequence completeness and E-value < 1E-5 (Table 4.6).

Table 4.6: Gene families of 32 *ail* homologs in *Yersinia* together with the BLASTP output where functional *ail* from *Y. pestis* CO92 was used as reference. The functional *ail* genes of human pathogenic species are highlighted in red.

Gene family	Subject	Subject locus tag	Subject length	Query coverage (%)	Query Identity (%)
1	*Y. similis* 228	CP007230.1_CDS_3282	184	100.00	42.70
	Y. pseudotuberculosis IP31758	CP000720.1_CDS_2171	183	98.91	45.36
		CP000720.1_CDS_1869	179	97.21	44.94
		CP000720.1_CDS_1114	179	100.00	99.44
	Y. pseudotuberculosis IP32953	BX936398.1_CDS_2930	179	100.00	99.44
		BX936398.1_CDS_1757	183	98.91	45.36
		BX936398.1_CDS_2167	179	97.21	44.94
	Y. pestis CO92	AL590842.1_CDS_1847	183	98.91	45.36
		AL590842.1_CDS_2879	179	100.00	100.00
		AL590842.1_CDS_2180	179	97.21	44.94
	Y. pestis KIM10+	AE009952.1_CDS_1968	179	97.21	44.94
		AE009952.1_CDS_2392	183	98.91	45.36
		AE009952.1_CDS_1299	179	100.00	100.00
2	*Y. enterocolitica* Y11	FR729477.2_CDS_21	178	100.00	74.30
	Y. enterocolitica 8081	AM286415.1_CDS_1784	178	100.00	73.74
	Y. similis 228	CP007230.1_CDS_63	179	100.00	91.62
		CP007230.1_CDS_1815	178	100.00	60.89

Table 4.6: Gene families of 32 *ail* homologs in *Yersinia* together with the BLASTP output where functional *ail* from *Y. pestis* CO92 was used as reference. The functional *ail* genes of human pathogenic species are highlighted in red, continued.

Gene family	Subject	Subject locus tag	Subject length	Query coverage (%)	Query Identity (%)
3	*Y. aldovae* 670-83	CP009781.1_CDS_3803	175	100.00	39.56
	Y. intermedia Y228	CP009801.1_CDS_3158	175	100.00	38.25
	Y. aleksiciae 159	CP011975.1_CDS_443	175	100.00	38.92
	Y. rohdei YRA	CP009787.1_CDS_3781	175	100.00	38.80
	Y. frederiksenii Y225	CP009364.1_CDS_591	175	100.00	38.92
	Y. kristensenii Y231	CP009997.1_CDS_1027	175	100.00	38.92
	Y. enterocolitica Y11	FR729477.2_CDS_1664	175	100.00	39.46
	Y. enterocolitica 8081	AM286415.1_CDS_2785	175	100.00	39.46
	Y. similis 228	CP007230.1_CDS_4080	174	100.00	40.66
	Y. pseudotuberculosis IP31758	CP000720.1_CDS_1436	174	100.00	40.66
	Y. pseudotuberculosis IP32953	BX936398.1_CDS_2607	174	100.00	40.66
	Y. pestis CO92	AL590842.1_CDS_2487	174	100.00	40.66
	Y. pestis KIM10+	AE009952.1_CDS_1634	174	100.00	40.66
	Y. ruckeri YRB	CP009539.1_CDS_3022	174	100.00	39.13
	Y. ruckeri Big Creek 74	CP011078.1_CDS_3141	174	100.00	38.46

I found that gene family 3 comprised of *ail* homologs which were present in each *Yersinia* species. All of the *ail* homologs in this family had only approximately 39% sequence identity to the reference *ail* protein sequence, suggesting that they might not be functionally similar to it. In the gene families 1 and 2, the *ail* homologs were only present in the human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*, but to the exception of non-pathogenic *Y. similis*. CP007230.1_CDS_63, one of the *ail* homologs in *Y. similis*, had very high sequence identity (91.6%) to the functional *ail* protein sequence, suggesting that they might be functionally similar to each other. I also found that the functional *ail* genes of *Y. pseudotuberculosis*-*Y. pestis* strains were grouped together with their respective *ail* homologs, proposing that these genes are paralogous to each other and the *ail* in these human pathogenic species might have arisen from gene duplication events.

To identify whether the *ail* gene arose from gene duplication events, I searched the sequences of the functional *ail* in pathogenic *Y. enterocolitica* against other *ail* homologs using BLASTP. (Table 4.7).

Table 4.7: BLASTP output of the functional *ail* from *Y. enterocolitica* 8081,which was used as query to search against *ail* homologs in *Yersinia*. Phylogroup-P species, which are highlighted in red, were in the top significant hits. The functional *ail* genes in pathogenic species are in bold.

Genome name	Subject ID	Identity (%)	E-value	Bit score
Y. similis 228	CP007230.1_CDS_63	74.86	3.72E-97	269
***Y. pseudotuberculosis* IP32953**	**BX936398.1_CDS_2930**	**73.74**	**1.10E-93**	**261**
***Y. pseudotuberculosis* IP31758**	**CP000720.1_CDS_1114**	**73.74**	**1.10E-93**	**261**
Y. similis 228	CP007230.1_CDS_1815	66.85	1.09E-90	253
Y. pseudotuberculosis IP31758	CP000720.1_CDS_2171	48.90	9.26E-56	164
Y. pseudotuberculosis IP32953	BX936398.1_CDS_1757	46.74	9.67E-56	164
Y. similis 228	CP007230.1_CDS_3282	46.49	2.12E-55	164
Y. pseudotuberculosis IP32953	BX936398.1_CDS_2167	46.39	2.95E-51	153
Y. pseudotuberculosis IP31758	CP000720.1_CDS_1869	46.39	2.95E-51	153
Y. ruckeri Big Creek 74	CP011078.1_CDS_3141	40.33	2.23E-42	130
Y. ruckeri YRB	CP009539.1_CDS_3022	40.33	7.85E-42	129
Y. aldovae 670-83	CP009781.1_CDS_3803	38.46	8.50E-39	121
Y. similis 228	CP007230.1_CDS_4080	38.30	9.44E-38	118
Y. pseudotuberculosis IP32953	BX936398.1_CDS_2607	38.30	1.57E-37	118
Y. pseudotuberculosis IP31758	CP000720.1_CDS_1436	38.30	1.57E-37	118
Y. intermedia Y228	CP009801.1_CDS_3158	37.36	1.80E-36	115
Y. kristensenii Y231	CP009997.1_CDS_1027	37.91	2.33E-36	115
Y. frederiksenii Y225	CP009364.1_CDS_591	37.91	2.33E-36	115
Y. enterocolitica Y11	FR729477.2_CDS_1664	36.81	2.41E-36	115
Y. enterocolitica 8081	AM286415.1_CDS_2785	36.81	2.41E-36	115
Y. rohdei YRA	CP009787.1_CDS_3781	37.16	7.10E-36	114
Y. aleksiciae 159	CP011975.1_CDS_443	36.81	5.47E-35	111

From the BLASTP output, phylogroup-P species were in the top significant hits. I also performed a pairwise comparison of *ail* and *ail* homologs between *Y. enterocolitica* strains and *Y. pseudotuberculosis* IP32953 (reference of phylogroup-P). The results are shown in Figure 4.7.

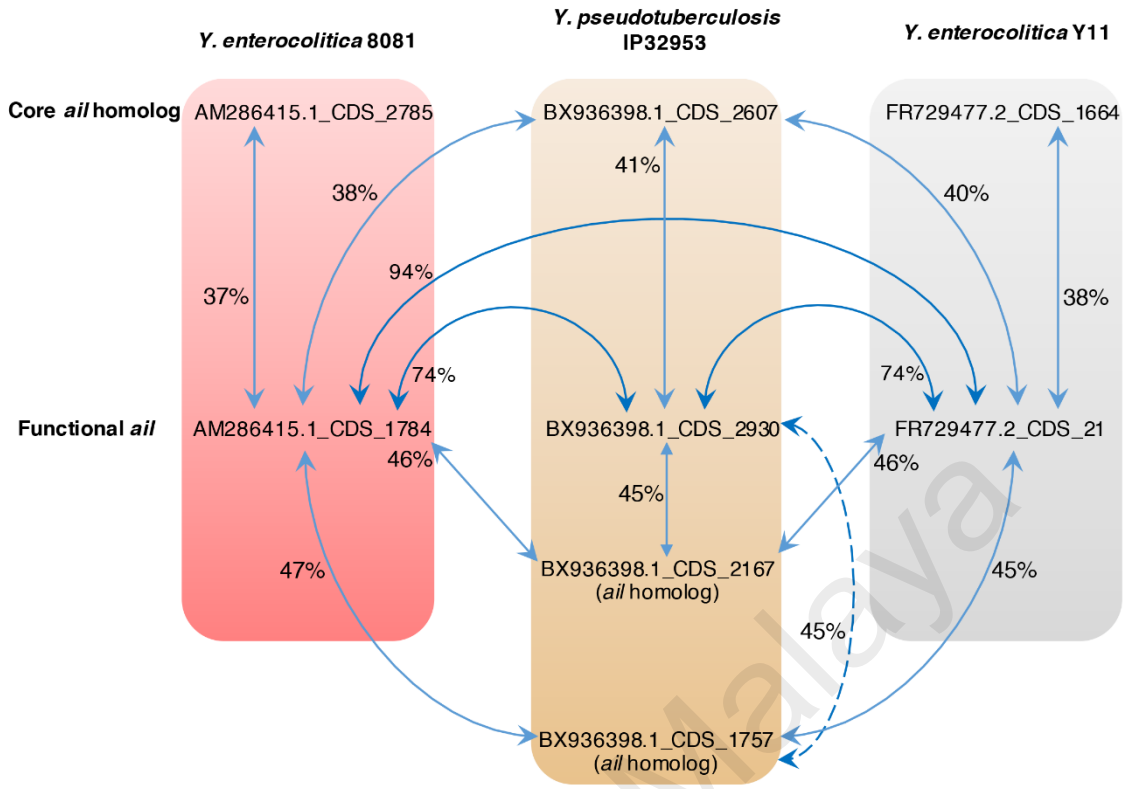

Figure 4.7: Pairwise percentage of identity between *ail* and *ail* homologs protein sequences for *Y. pseudotuberculosis* IP32953, *Y. enterocolitica* 8081 and Y11. Pairwise comparisons are indicated by blue double arrow pointing to two locus tags while the percentage of identity is labelled next to the arrow.

Although human pathogenic *Y. enterocolitica* had one functional *ail* and one *ail* homolog, I found that its functional *ail* was closer to the *ail* and *ail* homologs from phylogroup-P species compared to its own *ail* homolog. For instance, in Figure 4.7, the *ail* (AM286415.1_CDS_1784) from *Y. enterocolitica* 8081 had only 37% to its own *ail* homolog (AM286415.1_CDS_2785), but it had approximately 46% to *ail* paralogs (BX936398.1_CDS_2167 and BX936398.1_CDS_1757) and 74% to functional *ail* (BX936398.1_CDS_2930) from *Y. pseudotuberculosis* IP32953. *ail* from *Y. enterocolitica* Y11 also exhibited the same outcome. Taken all together, I hypothesize that the *ail* of *Y. enterocolitica* might be originated from the *ail* of *Y. pseudotuberculosis*, most probably due to lateral gene transfer. This hypothesis is made on the basis that top hits returned by the BLAST program could be used to predict the donor of laterally transferred gene (Ravenhall et al., 2015).

4.10 Genes exclusive to human pathogenic *Yersinia*

It would be important to identify which gene is exclusive to human pathogenic *Yersinia* species from different phylogroups to see if there is any convergent evolution. I found there were only a few genes which were present in both *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis* (Table 4.8).

Table 4.8: Genes exclusive to human pathogenic *Yersinia* from different phylogroups.

Gene locus	Gene location	Genome
ysc-yop T3SS	pYV plasmid	*Y. enterocolitica* *Y. pseudotuberculosis* *Y. pestis*
*yad*A	pYV plasmid	*Y. enterocolitica* *Y. pseudotuberculosis*
ybt	Chromosome	*Y. enterocolitica* 8081 *Y. pseudotuberculosis* IP32953 *Y. pestis*

The Ysc-Yop T3SS is a virulence factor that allows human pathogenic *Yersinia* to take over host cell signalling system and escape phagocytosis while YadA is involved in adhesion, serum resistance and Yop delivery (Cornelis, 2002a; Mikula et al., 2012). On the other hand, *ybt* locus is only present highly pathogenic *Yersinia*, such as *Y. enterocolitica* 8081 but not *Y. enterocolitica* Y11 (Pelludat et al., 1998). The locus is involved in yersiniabactin (a type of siderophore) synthesis and transport and it allows pathogenic *Yersinia* to scavenge iron in iron-limited environment. I did not find any metabolism gene unique to human pathogenic *Yersinia*, suggesting *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis* do not converge to utilize the same metabolic compound. Taken all together, pYV virulence plasmid is the most important virulence determinant ever acquired by human pathogenic *Yersinia* species.

4.11 Clustered Regularly-interspaced Short Palindromic Repeats in *Yersinia*

The CRISPR-Cas system is known to be a defence mechanism for bacteria to become immune to phage and plasmid (Haft et al., 2005). Spacers located in CRISPR array can provide resistance to foreign DNA if there is sequence homology between them. I found that only *Y. rohdei*, *Y. frederiksenii*, *Y. kristensenii*, *Y. similis* and human pathogenic *Y. pseudotuberculosis*-*Y. pestis* have the CRISPR-Cas and spacers. Another human pathogenic *Yersinia*, *Y. enterocolitica*, have lost both CRISPR-Cas system and spacers.

To identity the donors of these spacers, I performed BLASTN searches of the spacer sequences that were present in *Yersinia* genomes against the NCBI plasmid database (Geer et al., 2010). The BLASTN outputs are tabulated in Appendix D while the list of possible donors is summarized in Table 4.9.

Table 4.9: Summary of BLASTN outputs showing the possible donor of spacers found in *Yersinia* genomes. pYV virulence plasmid and pYE854 conjugative plasmid are in red text.

Genome and spacer locus	Possible donors of spacers
Y. frederiksenii Y225 CP009364.1_1_1745584 CP009364.1_2_1756249	*Burkholderia gladioli* BSR3 plasmid bgla_4p *Klebsiella pneumoniae* JM45 plasmid p1 *Lactococcus lactis* A76 plasmid pQA554 *Salmonella enterica* CFSAN000189 plasmid *Yersinia enterocolitica* plasmid pYE854
Y. kristensenii Y231 CP009997.1_1_26717 CP009997.1_2_37383	*Burkholderia gladioli* BSR3 plasmid bgla_4p *Klebsiella pneumoniae* BK31551 plasmid pBK31551 *Salmonella enterica* CFSAN000189 plasmid *Serratia marcescens* plasmid R830b *Yersinia enterocolitica* plasmid pYE854
Y. pestis CO92 AL590842.1_1_1773715	*Escherichia coli* plasmid pEC14_35 *Salmonella enterica* CFSAN001921 plasmid unnamed *Vibrio fischeri* MJ11 plasmid pMJ100 *Yersinia enterocolitica* 8081 plasmid pYVe8081 *Yersinia enterocolitica* Y11 plasmid pYVO3 *Yersinia enterocolitica* W22703 plasmid pYVe227
Y. pestis KIM10+ AE009952.1_3_2875781	*Escherichia coli* plasmid pEC14_35 *Salmonella enterica* CFSAN001921 plasmid unnamed *Vibrio fischeri* MJ11 plasmid pMJ100 *Yersinia enterocolitica* 8081 plasmid pYVe8081 *Yersinia enterocolitica* Y11 plasmid pYVO3 *Yersinia enterocolitica* W22703 plasmid pYVe227
Y. pseudotuberculosis IP32953 BX936398.1_2_2964849	*Enterobacter asburiae* LF7a plasmid pENTAS01 *Escherichia coli* E24377A plasmid pETEC_35 *Yersinia enterocolitica* plasmid pYV-WA314 *Yersinia enterocolitica* plasmid pYVa127/90 *Yersinia enterocolitica* Y11 plasmid pYVO3 *Yersinia enterocolitica* W22703 plasmid pYVe227
Y. similis 228 CP007230.1_2_4570273	*Bacillus cereus* FRI-35 plasmid p01 *Clostridium botulinum* D str. 1873 plasmid pCLG1 *Yersinia enterocolitica* 8081 plasmid pYVe8081 *Yersinia enterocolitica* Y11 plasmid pYVO3 *Yersinia pestis* CO92 plasmid pCD1 *Yersinia pseudotuberculosis* IP32953 pYV plasmid

I found that the donors to these spacers included pYV virulence plasmids (although they were not in the first top hit produced by BLAST; see Appendix D), pYE854 plasmid and plasmids from other genus. pYV plasmid is only found in human pathogenic *Yersinia* species and it encodes Ysc-Yop T3SS while pYE854 has been demonstrated to be able to mobilize pYV plasmid (Hammerl et al., 2008). I also found that spacers from different *Yersinia* phylogroups might able to target different plasmids. For instance, spacers found in *Y. frederiksenii* and *Y. kristensenii*, which were from phylogroup-E, could only recognize pYE854 plasmid; spacers found in phylogroup-P species could only recognize pYV plasmid. To be more precise, spacers present in human pathogenic *Y. pseudotuberculosis*-*Y. pestis* could only recognize pYV plasmid harboured by *Y. enterocolitica* while spacers present in human non-pathogenic *Y. similis* could recognize pYV plasmid harboured by both *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*. These observations suggest:

- CRISPR-Cas system present in the *Y. similis* could prevent the gain of pYV virulence plasmid harboured by the human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis* by targeting the conserved regions in the pYV plasmid.

- Human pathogenic *Y. pseudotuberculosis*-*Y. pestis* are unlikely to maintain the pYV plasmid from *Y. enterocolitica* as their spacers could recognize and fragment the pYV plasmid, if it was transferred laterally to *Y. pseudotuberculosis*-*Y. pestis*.

CHAPTER 5: RESULTS (PART II): THE SUBSPECIES OF *YERSINIA ENTEROCOLITICA*

5.1 Properties of *Yersinia enterocolitica* genomes

A total of 73 *Y. enterocolitica* strains were used to study the subspecies of *Y. enterocolitica*. The summary of genome annotation of each strain is tabulated in Appendix C. Briefly. The median genome size, guanine-cytosine content and number of open reading frames in *Y. enterocolitica* were 4,594,630, 46.94% and 4,154, respectively.

5.2 Average nucleotide identity between *Yersinia enterocolitica* genomes

To study the genomic similarities between *Y. enterocolitica* strains, the ANI values of each genome pairs were calculated. I found that the lowest ANI value was 95.09%, confirming all of the *Y. enterocolitica* strains used in this study belonged to the same species because ANI value exceeded the threshold (95%) to be considered as single species (Konstantinidis & Tiedje, 2005).

5.3 Gene families of *Yersinia enterocolitica*

To study the gene families in *Y. enterocolitica*, the chromosomal protein sequences in all strains were clustered into 10,131 gene families based on 25% sequence identity, 50% sequence completeness and E-value of 1E-5. Percentage of orthologous, co-orthologous, dispensable and strain-specific gene families are shown in Figure 5.1.

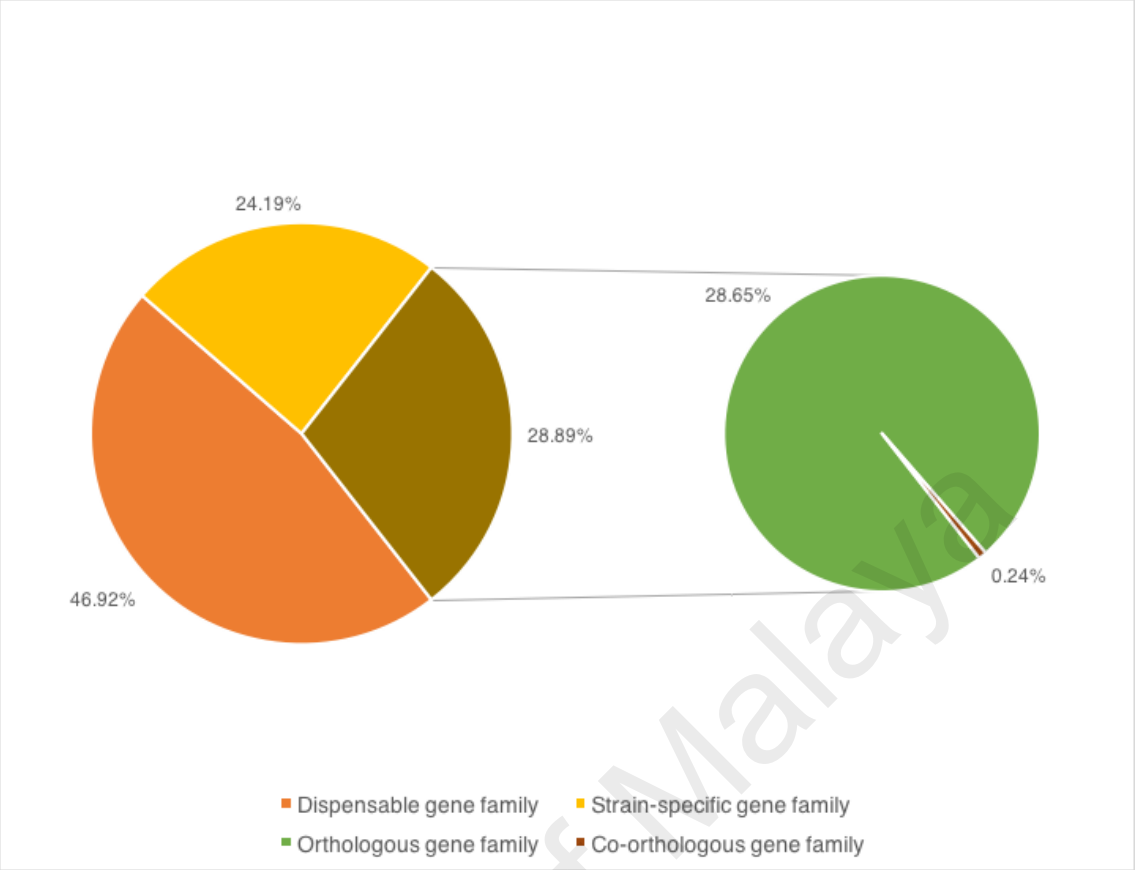

Figure 5.1: Percentage of orthologous, co-orthologous, dispensable and strain-specific gene families present in *Y. enterocolitica*.

I found that the orthologous and co-orthologous gene families contributed about a quarter (28.89%) to the total gene families of *Y. enterocolitica*. The proportion was still far less than combination of dispensable and strain-specific gene families (71.11%). These values suggest that the genomes of *Y. enterocolitica* are likely mosaic and this species might consist of heterogeneous collection of strains, showing consistency with a previous report (Segerman, 2012).

5.4 Phylogenetic relationships between *Yersinia enterocolitica* strains

5.4.1 *Yersinia enterocolitica* supermatrix tree

To infer the phylogenetic relationships between *Y. enterocolitica* strains, I first reconstructed a supermatrix tree based on non-recombinant super-sequence with 1,138,594 nucleotides present in all strains (Figure 5.2).

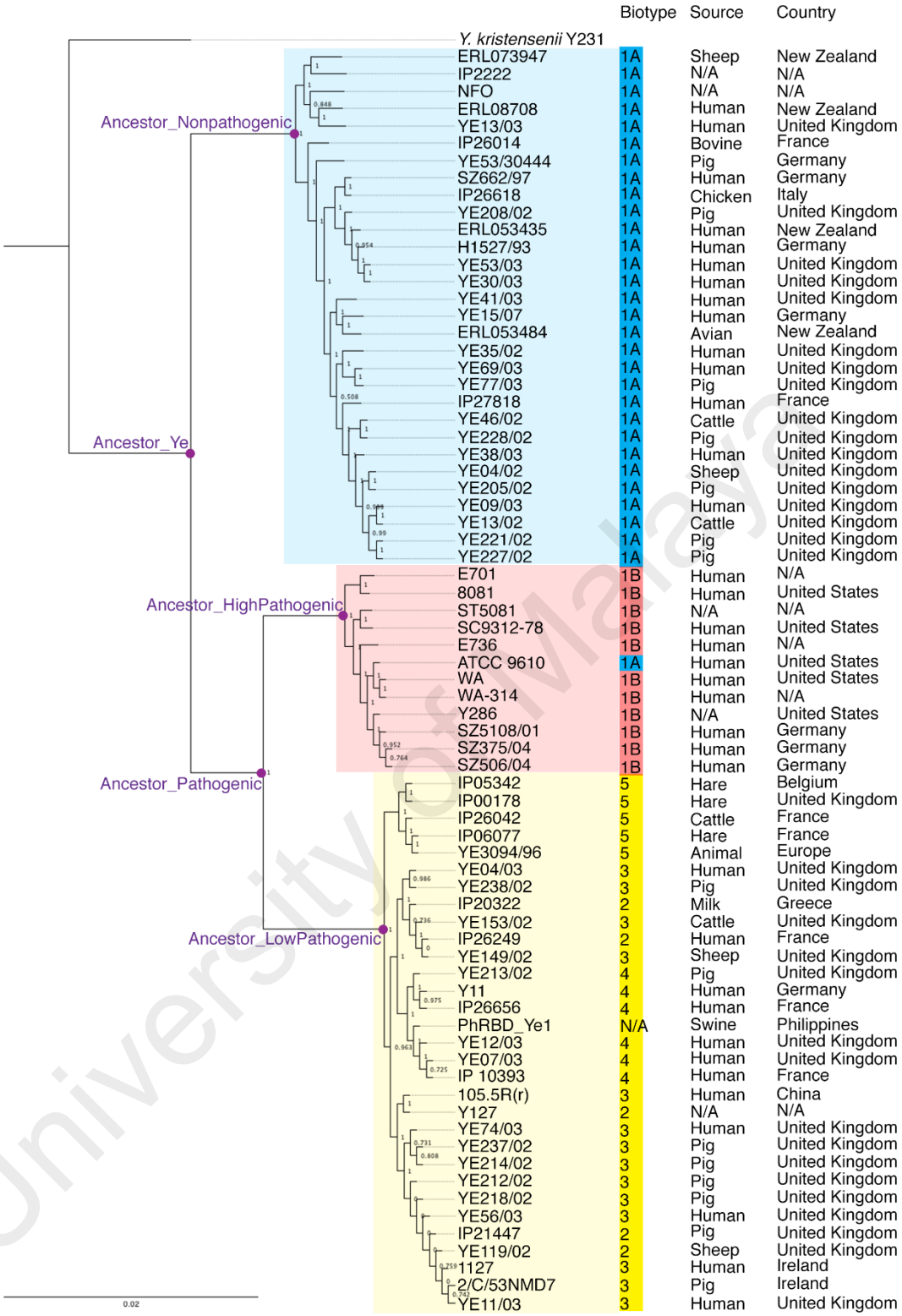

Figure 5.2: *Y. enterocolitica* supermatrix tree constructed from non-recombinant super-sequences and rooted by *Y. kristensenii* Y231. Biotype, isolation source and country are labelled next to the strain name. Non-pathogenic biotype 1A, low pathogenic biotype 2-5 and highly pathogenic biotype 1B are highlighted in cyan, yellow and magenta respectively. Non-pathogenic subspecies, low pathogenic subspecies and highly pathogenic subspecies are highlighted in cyan, yellow and magenta respectively. Ancestors of interest are labelled in violet text. Bootstrap values of internal nodes are shown.

From the supermatrix tree, I found that all of the *Y. enterocolitica* strains were resolved into three distinct phylogroups: (1) highly pathogenic phylogroups which consists of biotype 1B except ATCC 9610, which is non-pathogenic biotype 1A, (2) low pathogenic phylogroup which consists of biotypes 2 to 5, (3) non-pathogenic phylogroup which consists of biotype 1A. "The most recent ancestor of all *Y. enterocolitica* strains" (Ancestor_Ye) diverged into two: "the most recent ancestor of non-pathogenic *Y. enterocolitica* strain" (Ancestor_Nonpathogenic) and "the most recent ancestor of pathogenic *Y. enterocolitica* strain" (Ancestor_Pathogenic). Ancestor_Pathogenic further diverged into low pathogenic and highly pathogenic phylogroups. To ease my explanations that followed, I designated Ancestor_LowPathogenic and Ancestor_HighPathogenic to be the last common ancestor for low pathogenic and highly pathogenic phylogroups respectively. I also found that a higher number of nucleotide substitutions had occurred in low pathogenic phylogroup, as indicated by a longer branch from its ancestor (Ancestor_LowPathogenic) to Ancestor_Ye. In comparison, the respective branches from Ancestor_Nonpathogenic and Ancestor_HighPathogenic to Ancestor_Ye were shorter. My data also suggest that there might be three subspecies exist in the *Y. enterocolitica*.

5.4.2 *Yersinia enterocolitica* gene content-based phylogenetic tree

Based on the information of the presence and absence of gene families in each genome, a gene content phylogenetic tree was reconstructed to infer the phylogenetic relationship of *Y. enterocolitica* strains in Figure 5.3.

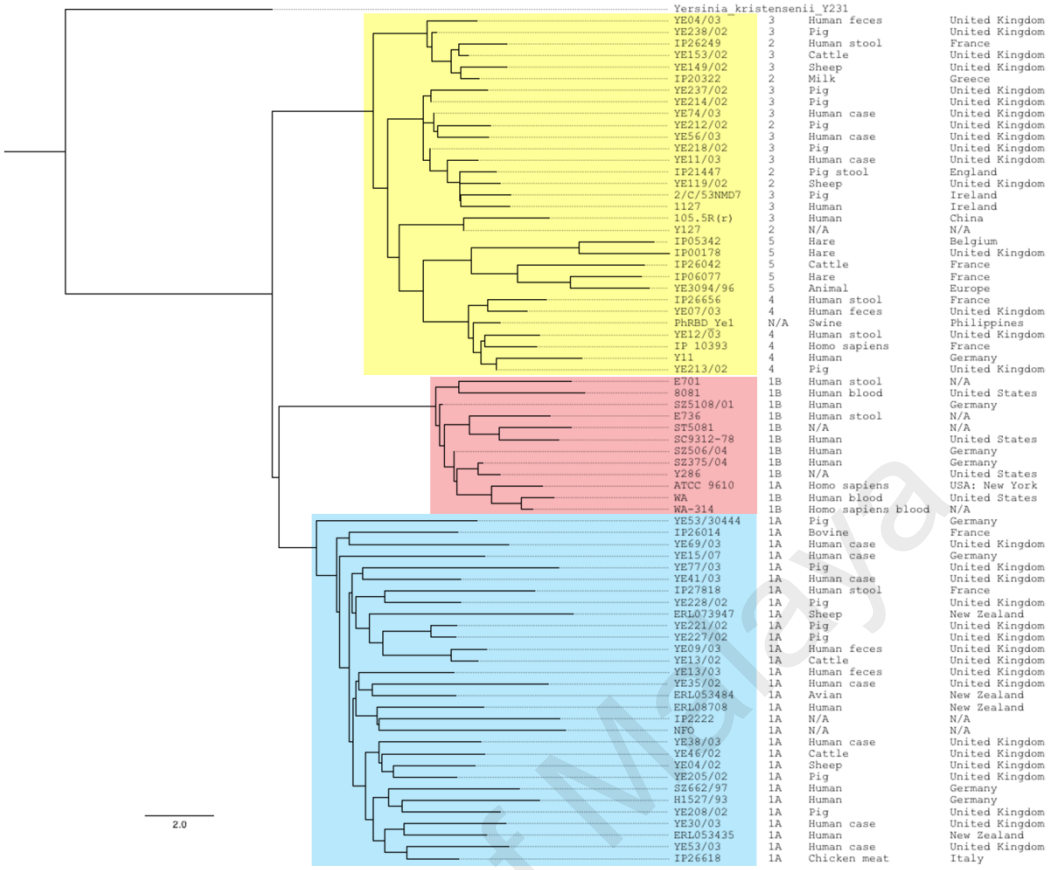

Figure 5.3: *Y. enterocolitica* gene content-based phylogenetic tree constructed based on presence and absence of gene family in each genome and rooted by *Y. kristensenii* Y231. The tree exhibits similar phyletic patterns with supermatrix tree. Highly pathogenic, low pathogenic and non-pathogenic phylogroups are highlighted in magenta, yellow and cyan respectively.

I found that the *Y. enterocolitica* gene content phylogenetic tree exhibited similar topology with its supermatrix tree, whereby all strains were grouped into highly pathogenic, low pathogenic and non-pathogenic biogroups. This suggests that each phylogroup has its own unique set of gene families which are not found in the rest.

5.5 Phylogenetic network and recombination in *Yersinia enterocolitica*

To study the recombination events in the *Y. enterocolitica*, I first reconstructed a phylogenetic network to display conflicting phylogenetic signals within *Y. enterocolitica* (Figure 5.4).

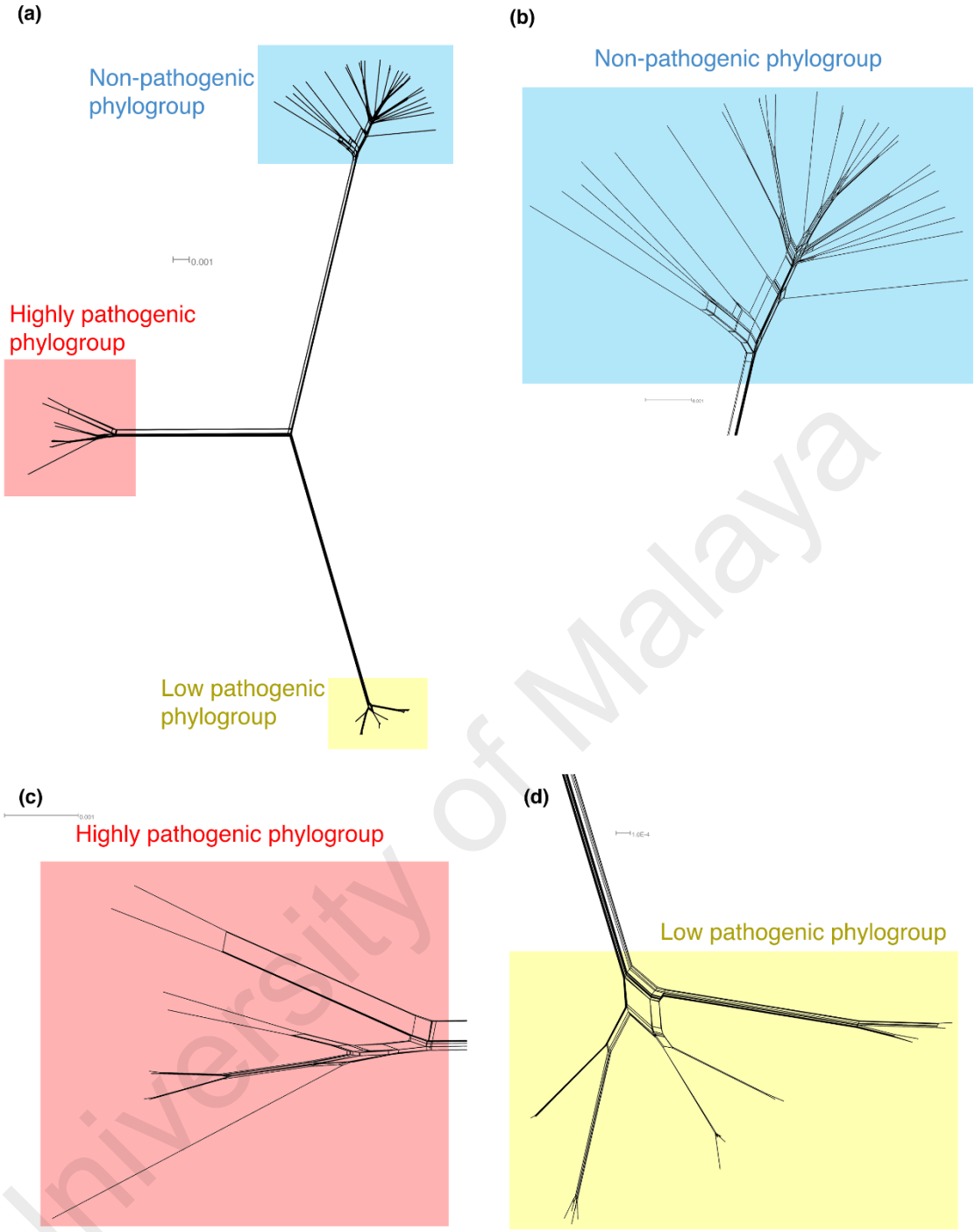

Figure 5.4: Phylogenetic network of *Y. enterocolitica* strains constructed using non-recombinant super-sequences. (a) Phylogenetic network of *Y. enterocolitica* shows conflicting phylogenetic signals between strains and demarcates all strains into three phylogroups: highly pathogenic, low pathogenic and non-pathogenic phylogroups, which are highlighted by magenta, yellow and cyan respectively. (b) Zoomed reticulation of non-pathogenic phylogroup. (c) Zoomed reticulation of highly pathogenic phylogroup. (d) Zoomed reticulation of low pathogenic phylogroup.

I found that the phylogenetic network clearly exhibited three phylogroups, phyletic patterns which were similar to *Y. enterocolitica* supermatrix tree. These three phylogroups, highly pathogenic, low pathogenic and non-pathogenic phylogroups, were connected by long edges without much conflicting signals. However, by zooming into each phylogroup, I found that there were reticulations appeared near to the tips of the phylogenetic network. Thus, reticulation or recombination mainly occurred within strains belonged to the same phylogroup, rather than spanned across phylogroups. This suggests that the intra-phylogroup recombination is likely to occur more often than the inter-phylogroup recombination.

Besides the visual inspections, I also estimated the rate of recombination and mutations of all *Y. enterocolitica* strains and other strains belonged to the same phylogroup to obtain statistical support (Table 5.1).

Table 5.1: Estimation of the rate of recombination and mutation in three different *Y. enterocolitica* dataset.

Dataset	Rate of recombination to mutation R/θ	Mean DNA import length δ	Mean divergence of imported DNA ν	Ratio of mutation to recombination M→R	Relative effect of recombination to mutation r/m
All *Y. enterocolitica* strains	0.2971	95.5219	0.0219	3.3659:1	0.6215
Highly pathogenic phylogroup	0.2301	104.2259	0.0820	4.3459:1	1.9666
Low pathogenic phylogroup	0.0682	45.7149	0.2574	14.6628:1	0.8025
Nonpathogenic phylogroup	0.7915	169.5374	0.02316	1.2634:1	3.1078

In general, I found that the rate of recombination to mutation was lower than 1 for all four datasets, indicating that mutations played a major role in shaping the genome of *Y. enterocolitica*. When all strains were taken into consideration, mutation was estimated to have occurred three times more than the recombination. If each phylogroup was analysed independently, I found that the low pathogenic phylogroup had the least recombination events as indicated by its lowest R/θ and highest M→R values (0.0682 and 14.6628:1 respectively) compared to the highly pathogenic (0.2301 and 4.3459:1 respectively) and non-pathogenic phylogroups (0.7915 and 1.2634:1 respectively). On the other hand, the non-pathogenic phylogroup had the most recombination events as indicated by its highest R/θ and lowest M→R values (0.7915 and 1.2634:1 respectively) compared to the other two phylogroups. Overall, although mutations tend to be the dominant force in *Y. enterocolitica*, their rates vary among phylogroups.

5.6 Gene gain-and-loss in *Yersinia enterocolitica*

To study the ancestral gene gain-and-loss events before emergence of three *Y. enterocolitica* phylogroups, I had performed gene gain-and-loss analysis. In the following subsections, I discussed the acquired and lost genes in hypothetical ancestors of interest, which are Ancestor_Ye, Ancestor_Nonpathogenic, Ancestor_Pathogenic, Ancestor_HighPathogenic and Ancestor_LowPathogenic (see Figure 5.2).

5.6.1 Emergence of the most recent ancestor of all *Yersinia enterocolitica* strains (Ancestor_Ye)

Ancestor_Ye was the most recent ancestor shared by highly pathogenic, low pathogenic and non-pathogenic phylogroups. I found that it had gained several virulence genes which are important for pathogenesis of human pathogenic *Yersinia*, such as *myf*ABCEF and *inv* genes (Mikula et al., 2012). This suggests Ancestor_Ye could be pathogenic.

On the other hand, Ancestor_Ye had lost CRISPR-Cas system, which is proposed as one of the defence mechanisms used by bacteria to resist foreign DNA materials, including phage and plasmid (Marraffini & Sontheimer, 2008). However, I found that the system was present in *Y. kristensenii* Y231 and the spacers had sequence similarity to pYE854 conjugative plasmid, which is able to mobilize pYV virulence plasmid (Hammerl et al., 2008). Hence, the loss of CRISPR-Cas system in *Y. enterocolitica* seems important to acquire the pYV plasmid.

5.6.2 Emergence of the most recent ancestor of non-pathogenic *Yersinia enterocolitica* strains (Ancestor_Nonpathogenic)

Ancestor_Nonpathogenic was the direct descendant of Ancestor_Ye and it led to the emergence of non-pathogenic phylogroup. I found that Ancestor_Apathogenic gained D-serine metabolism genes (*dsd*ACX). A recent study showed that ability to degrade D-serine could affect the niche selection of different types of *Escherichia coli* strains and enables uropathogenic *E. coli* to have survival advantage in the urinary tract which has abundant D-serine (Connolly et al., 2015). Other genes gained by Ancestor_Apathogenic are the L-fucose metabolism genes (*fuc*OAPIKUR). The utilization of L-fucose provides competitive advantage and is associated with the virulence lifestyle of *Campylobacter jejuni*, a gastrointestinal pathogen (Choi et al., 2009). Thus, acquisition of *dsd* and *fuc* loci could open new niches and beneficial for non-pathogenic *Y. enterocolitica* strains.

5.6.3 Emergence of the most recent ancestor of pathogenic *Yersinia enterocolitica* strains (Ancestor_Pathogenic)

Similar with Ancestor_Nonpathogenic, Ancestor_Pathogenic was also the direct ancestor of Ancestor_Ye and it led to the emergence of highly pathogenic and low pathogenic phylogroups. I found that Ancestor_Pathogenic gained *ail*, a known virulence factor that

is present only in human pathogenic *Yersinia* (Mikula et al., 2012). The gene is involved in adherence and invasion into host cell. Thus, the acquisition of *ail* might be one of the key factors for the pathogenic *Y. enterocolitica* to arise.

5.6.4 Emergence of the most recent ancestor of low pathogenic *Yersinia enterocolitica* strains (Ancestor_LowPathogenic)

One of the direct descendants of Ancestor_Pathogenic was Ancestor_LowPathogenic, which led to the emergence of low pathogenic phylogroup. I found that Ancestor_LowPathogenic had gained ABC transporter genes for dispersin (*aat*PABCD). Previous study proposed that *aat* locus was important to the pathogenesis of enteroaggregative *E. coli* (Nishi et al., 2003). Hence, the *aat* locus might confer virulence traits expressed by low pathogenic strains. Ancestor_LowPathogenic also acquired a pair of toxin-antitoxin (TA) genes, namely *mqs*R and *mqs*A. Both genes were reported to be the most highly up-regulated gene in persistent *E. coli* cells where it regulates other physiological genes (Brown et al., 2009). Thus, the *mqs*R-*mqs*A TA gene might be another important gene for low pathogenic strains to overcome stress from host immune system.

5.6.5 Emergence of the most recent ancestor of highly pathogenic *Yersinia enterocolitica* strains (Ancestor_HighPathogenic)

Similar to Ancestor_LowPathogenic, Ancestor_HighPathogenic was also the direct descendant of Ancestor_Pathogenic, which led to the emergence of highly pathogenic phylogroup. I found that Ancestor_HighPathogenic had gained many virulence genes, including *yts*1 locus (encodes Type Two Secretion System [T2SS]), *ybt* locus (yersiniabactin synthesis and transport genes) and *ysa* locus (chromosomal-encoded

T3SS). All of these genes are known to be hallmarks of highly pathogenic *Y. enterocolitica* strains (Foultier et al., 2002; Iwobi et al., 2003; Pelludat et al., 1998).

On the other hand, another chromosomal-encoded T3SS was lost in Ancestor_HighPathogenic. The lost T3SS was different from T3SS encoded in the *ysa* locus. Instead, it was more similar to T3SS encoded by "*Salmonella* pathogenicity island 2" of *Salmonella typhimurium* (Batzilla et al., 2011a). The function of this T3SS is unknown in *Yersinia*, but I hypothesize that the loss of this chromosomal T3SS might be in exchange with *ysa*-T3SS.

5.6.6 Emergence of non-pathogenic *Yersinia enterocolitica* ATCC 9610 in the highly pathogenic phylogroup

From the supermatrix tree, I found that ATCC 9610 was the only non-pathogenic *Y. enterocolitica* strain which was grouped together with the highly pathogenic strains such as *Y. enterocolitica* 8081 and WA-314 (Garzetti et al., 2012). Hence, the gene gain-and-loss events might have played important roles in the evolution of *Y. enterocolitica* ATCC 9610. Despite being a non-pathogenic strain (Neubauer et al., 2000), I found that the *Y. enterocolitica* ATCC 9610 strain had only lost chromosome-borne *ail*, and the pYV plasmid-borne *yad*A and *ysc-yop* T3SS. The hallmarks of highly pathogenic *Y. enterocolitica* such as *ybt*, *ysa*-T3SS and *yts*1-T2SS were still present in the genome of ATCC 9610 (Foultier et al., 2002; Iwobi et al., 2003; Pelludat et al., 1998). These observations suggest that the loss of *ail* and pYV virulence plasmid might be sufficient to render the *Y. enterocolitica* ATCC 9610 to be harmless to human.

5.7 *inv* homologs in *Yersinia enterocolitica*

Gene gain-and-loss analysis suggests that Ancestor_Ye had gained *inv*. I found no *inv* or *inv* homolog was lost in non-pathogenic *Y. enterocolitica* strains. To further validate the presence of *inv* or its homolog in these non-pathogenic strains, I performed BLASTP searches of the functional *inv* protein sequence (835 amino acids) of highly pathogenic *Y. enterocolitica* 8081 against all protein sequences of non-pathogenic strains and the results are summarized in Table 5.2.

Table 5.2: BLASTP outputs showing high identity and high sequence coverage between the functional *inv* of highly pathogenic *Y. enterocolitica* 8081 and *inv* homologs of non-pathogenic *Y. enterocolitica* strains.

Strain name	Subject accession	Query identity (%)	Query start	Query end	Subject start	Subject end	Subject length	Query coverage (%)	Subject coverage (%)
YE46/02	CTKK01000008.1_CDS_30	89.95	1	835	1	836	836	100.00	100.00
YE69/03	CTRB01000011.1_CDS_105	89.71	1	835	1	836	836	100.00	100.00
YE04/02	CTKX01000009.1_CDS_31	89.59	1	835	1	836	836	100.00	100.00
YE205/02	CTIV01000010.1_CDS_106	89.59	1	835	1	836	836	100.00	100.00
H1527/93	CQCE01000010.1_CDS_35	89.47	1	835	1	836	836	100.00	100.00
SZ662/97	CGBV01000010.1_CDS_104	89.47	1	835	1	836	836	100.00	100.00
YE30/03	CTEV01000011.1_CDS_29	89.47	1	835	1	836	836	100.00	100.00
YE53/03	HF571988.1_CDS_2747	89.47	1	835	1	836	836	100.00	100.00
YE77/03	CTKV01000013.1_CDS_35	89.47	1	835	1	836	836	100.00	100.00
IP26014	CQAE01000010.1_CDS_35	89.35	1	835	1	836	836	100.00	100.00
YE09/03	CTKQ01000009.1_CDS_30	89.35	1	835	1	836	836	100.00	100.00
YE13/02	CTIZ01000010.1_CDS_30	89.35	1	835	1	836	836	100.00	100.00
YE221/02	CTIP01000011.1_CDS_40	89.35	1	835	1	836	836	100.00	100.00
YE227/02	CTJE01000011.1_CDS_106	89.35	1	835	1	836	836	100.00	100.00
YE38/03	CTKL01000012.1_CDS_35	89.35	1	835	1	836	836	100.00	100.00
YE208/02	CTIT01000001.1_CDS_35	89.23	1	835	1	836	836	100.00	100.00
IP26618	CPYU01000011.1_CDS_104	87.93	1	835	1	837	837	100.00	100.00
YE228/02	CTRG01000012.1_CDS_42	85.68	1	835	1	837	837	100.00	100.00
ERL073947	CFLA01000002.1_CDS_272	85.56	1	835	1	837	837	100.00	100.00
ERL053435	CQAJ01000009.1_CDS_35	85.44	1	835	1	837	837	100.00	100.00
IP27818	CPZF01000009.1_CDS_117	85.32	1	835	1	837	837	100.00	100.00

Table 5.2: BLASTP outputs showing high identity and high sequence coverage between the functional *inv* of highly pathogenic *Y. enterocolitica* 8081 and *inv* homologs of non-pathogenic *Y. enterocolitica* strains, continued.

Strain name	Subject accession	Query identity (%)	Query start	Query end	Subject start	Subject end	Subject length	Query coverage (%)	Subject coverage (%)
YE53/30444	CQEI01000015.1_CDS_76	85.32	1	835	1	837	837	100.00	100.00
YE13/03	CTIU01000016.1_CDS_65	85.20	1	835	1	837	837	100.00	100.00
YE35/02	CTKN01000013.1_CDS_38	85.20	1	835	1	837	837	100.00	100.00
ERL08708	CPZT01000013.1_CDS_106	85.19	56	835	1	782	782	93.41	100.00
ERL053484	CWGL01000008.1_CDS_36	85.08	1	835	1	837	837	100.00	100.00
YE15/07	CPYS01000013.1_CDS_35	85.08	1	835	1	837	837	100.00	100.00
YE41/03	CTIN01000009.1_CDS_39	85.08	1	835	1	837	837	100.00	100.00
NFO	CACY01000058.1_CDS_35	84.84	1	835	1	837	837	100.00	100.00
IP2222	CACZ01000022.1_CDS_109	84.61	1	835	1	837	837	100.00	100.00
Minimum		84.61					782	93.41	100
Maximum		89.95					837	100	100
Median		89.29					836	100	100

From the BLASTP search results, I found that the *inv* homologs were present in non-pathogenic *Y. enterocolitica* strains and had high sequence identity (more than 84%) and high sequence coverage (median = 100%) to the protein sequence of functional *inv* in highly pathogenic *Y. enterocolitica* 8081. Furthermore, the alignments were perfect because all of them (except ERL08708) started from first amino acids and ended at last amino acids in both query and subject sequences. These observations suggest that the *inv* homologs of non-pathogenic strains might be functionally similar to the known *inv* gene of the pathogenic *Y. enterocolitica*.

5.8 Pseudogenized *ail* virulence gene in non-pathogenic *Yersinia enterocolitica*

Based on the reconstruction of gene gain-and-loss, I found that *ail* virulence gene was acquired by Ancestor_Pathogenic, which is shared by both low and highly pathogenic *Y. enterocolitica* phylogroups. To examine whether *Y. enterocolitica* strains possess additional copies of genes homologous to the *ail* gene, I performed BLASTP and TBLASTN searches of the functional *ail* of highly pathogenic *Y. enterocolitica* 8081 against the protein and genome sequences of all *Y. enterocolitica* strains that I used in this study (Table 5.3).

Table 5.3: BLASTP outputs showing the presence of *ail* and *ail* homologs in *Y. enterocolitica* strains.

Strain name	Subject locus tag	Query sequence identity (%)	Query sequence coverage (%)	Subject sequence coverage (%)
Non-pathogenic *Y. enterocolitica* strains				
ERL073947	CFLA01000008.1_CDS_153	36.81	99.44	99.43
IP2222	CACZ01000021.1_CDS_49	36.81	99.44	99.43
NFO	CACY01000060.1_CDS_52	36.81	99.44	99.43
ERL08708	CPZT01000006.1_CDS_153	36.81	99.44	99.43
YE13/03	CTIU01000004.1_CDS_153	36.81	99.44	99.43
IP26014	CQAE01000006.1_CDS_153	36.81	99.44	99.43
YE53/30444	**CQEI01000020.1_CDS_43**	**73.68**	**51.69**	**96.91**
YE53/30444	**CQEI01000007.1_CDS_92**	**36.81**	**99.44**	**99.43**
SZ662/97	CGBV01000002.1_CDS_156	36.81	99.44	99.43
IP26618	CPYU01000002.1_CDS_156	36.81	99.44	99.43
YE208/02	CTIT01000001.1_CDS_295	36.81	99.44	99.43
ERL053435	CQAJ01000008.1_CDS_25	36.81	99.44	99.43
H1527/93	CQCE01000002.1_CDS_154	36.81	99.44	99.43
YE53/03	HF571988.1_CDS_3010	36.81	99.44	99.43
YE30/03	CTEV01000002.1_CDS_153	36.81	99.44	99.43
YE41/03	CTIN01000007.2_CDS_92	36.81	99.44	99.43
YE15/07	CPYS01000004.1_CDS_93	36.26	99.44	99.43
ERL053484	CWGL01000001.1_CDS_461	36.81	99.44	99.43
YE35/02	CTKN01000008.1_CDS_154	36.81	99.44	99.43
YE69/03	CTRB01000015.1_CDS_2	36.81	99.44	99.43
YE77/03	CTKV01000005.1_CDS_154	36.81	99.44	99.43
IP27818	CPZF01000007.1_CDS_155	36.81	99.44	99.43
YE46/02	CTKK01000005.1_CDS_153	36.81	99.44	99.43
YE228/02	CTRG01000007.1_CDS_153	36.81	99.44	99.43
YE38/03	CTKL01000006.1_CDS_94	36.26	99.44	99.43
YE04/02	CTKX01000006.1_CDS_155	36.81	99.44	99.43
YE205/02	CTIV01000009.1_CDS_155	36.81	99.44	99.43
YE09/03	CTKQ01000006.1_CDS_155	36.81	99.44	99.43
YE13/02	CTIZ01000006.1_CDS_92	36.81	99.44	99.43
YE221/02	CTIP01000007.1_CDS_93	36.81	99.44	99.43
YE227/02	CTJE01000006.1_CDS_155	36.81	99.44	99.43
Highly pathogenic *Y. enterocolitica* strains				
E701	CWIY01000070.1_CDS_2	100	99.44	99.44
E701	CWIY01000013.1_CDS_61	36.81	99.44	99.43
8081	AM286415.1_CDS_1784	100	99.44	99.44
8081	AM286415.1_CDS_2785	36.81	99.44	99.43

Table 5.3: BLASTP outputs showing the presence of *ail* and *ail* homologs in *Y. enterocolitica* strains, continued.

Strain name	Subject locus tag	Query sequence identity (%)	Query sequence coverage (%)	Subject sequence coverage (%)
ST5081	CWJB01000064.1_CDS_1	98.88	99.44	99.44
ST5081	CWJB01000020.1_CDS_54	36.81	99.44	99.43
SC9312-78	CQDJ01000060.1_CDS_7	98.88	99.44	99.44
SC9312-78	CQDJ01000021.1_CDS_54	36.81	99.44	99.43
E736	CWIZ01000084.1_CDS_2	98.31	99.44	99.44
E736	CWIZ01000027.1_CDS_53	36.81	99.44	99.43
ATCC 9610	**KN150735.1_CDS_744**	**36.81**	**99.44**	**99.43**
WA-314	AKKR01000003.1_CDS_1	98.88	99.44	99.44
WA-314	AKKR01000025.1_CDS_53	36.81	99.44	99.43
WA	CP009367.1_CDS_1224	98.88	99.44	99.44
WA	CP009367.1_CDS_2145	36.81	99.44	99.43
Y286	CPYO01000058.1_CDS_6	98.88	99.44	99.44
Y286	CPYO01000017.1_CDS_53	36.81	99.44	99.43
SZ5108/01	CPZL01000030.1_CDS_52	98.88	99.44	99.44
SZ5108/01	CPZL01000019.1_CDS_14	36.81	99.44	99.43
SZ375/04	CGBA01000067.1_CDS_1	98.88	99.44	99.44
SZ375/04	CGBA01000021.1_CDS_14	36.81	99.44	99.43
SZ506/04	CQCF01000067.1_CDS_2	98.88	99.44	99.44
SZ506/04	CQCF01000016.1_CDS_14	36.81	99.44	99.43
Low pathogenic *Y. enterocolitica* strains				
IP05342	CPXJ01000095.1_CDS_1	94.38	99.44	99.44
IP05342	CPXJ01000037.1_CDS_6	36.81	99.44	99.43
IP00178	CTFT01000107.1_CDS_1	94.38	99.44	99.44
IP00178	CTFT01000032.1_CDS_47	36.81	99.44	99.43
IP26042	CGGL01000090.1_CDS_1	89.84	99.44	99.47
IP26042	CGGL01000002.1_CDS_6	36.81	99.44	99.43
IP06077	CPYG01000089.1_CDS_1	89.84	99.44	99.47
IP06077	CPYG01000002.1_CDS_271	36.81	99.44	99.43
YE3094/96	HF933426.1_CDS_3300	89.84	99.44	99.47
YE3094/96	HF933426.1_CDS_1670	36.81	99.44	99.43
YE04/03	CTEQ01000042.1_CDS_2	94.38	99.44	99.44
YE04/03	CTEQ01000015.1_CDS_97	36.81	99.44	99.43
YE238/02	CTFS01000048.1_CDS_2	94.38	99.44	99.44
YE238/02	CTFS01000017.1_CDS_75	36.81	99.44	99.43
IP20322	CQBQ01000043.1_CDS_2	94.38	99.44	99.44
IP20322	CQBQ01000003.1_CDS_5	36.81	99.44	99.43
YE153/02	CTJG01000091.1_CDS_2	94.38	99.44	99.44

Table 5.3: BLASTP outputs showing the presence of *ail* and *ail* homologs in *Y. enterocolitica* strains, continued.

Strain name	Subject locus tag	Query sequence identity (%)	Query sequence coverage (%)	Subject sequence coverage (%)
YE153/02	CTJG01000013.1_CDS_5	36.81	99.44	99.43
IP26249	CGBR01000041.1_CDS_30	94.38	99.44	99.44
IP26249	CGBR01000001.1_CDS_5	36.81	99.44	99.43
YE149/02	HF933424.1_CDS_3570	94.38	99.44	99.44
YE149/02	HF933424.1_CDS_1965	36.81	99.44	99.43
YE213/02	CTEZ01000116.1_CDS_2	94.38	99.44	99.44
YE213/02	CTEZ01000008.1_CDS_98	36.81	99.44	99.43
Y11	FR729477.2_CDS_21	94.38	99.44	99.44
Y11	FR729477.2_CDS_1664	36.81	99.44	99.43
IP26656	CGBC01000085.1_CDS_2	94.38	99.44	99.44
IP26656	CGBC01000009.1_CDS_99	36.81	99.44	99.43
PhRBD_Ye1	AGQO01000064.1_CDS_2	94.38	99.44	99.44
PhRBD_Ye1	AGQO01000009.1_CDS_5	36.81	99.44	99.43
YE12/03	HF933425.1_CDS_2913	94.38	99.44	99.44
YE12/03	HF933425.1_CDS_1332	36.81	99.44	99.43
YE07/03	CTKR01000077.1_CDS_2	94.38	99.44	99.44
YE07/03	CTKR01000005.1_CDS_90	36.81	99.44	99.43
IP 10393	CAOV01000002.1_CDS_227	94.38	99.44	99.44
IP 10393	CAOV01000008.1_CDS_997	36.81	99.44	99.43
105.5R(r)	CP002246.1_CDS_3069	94.38	99.44	99.44
105.5R(r)	CP002246.1_CDS_1473	36.81	99.44	99.43
Y127	CWIU01000059.1_CDS_2	94.38	99.44	99.44
Y127	CWIU01000003.1_CDS_5	36.81	99.44	99.43
YE74/03	CWGM01000018.1_CDS_20	94.38	99.44	99.44
YE74/03	CWGM01000001.1_CDS_197	36.81	99.44	99.43
YE237/02	CTQV01000054.1_CDS_20	94.38	99.44	99.44
YE237/02	CTQV01000003.1_CDS_96	36.81	99.44	99.43
YE214/02	CTRD01000055.1_CDS_2	94.38	99.44	99.44
YE214/02	CTRD01000003.1_CDS_78	36.81	99.44	99.43
YE212/02	HF933206.1_CDS_1087	94.38	99.44	99.44
YE212/02	HF933206.1_CDS_2491	36.81	99.44	99.43
YE218/02	CTKS01000059.1_CDS_20	94.38	99.44	99.44
YE218/02	CTKS01000002.1_CDS_78	36.81	99.44	99.43
YE56/03	HF933423.1_CDS_1743	94.38	99.44	99.44
YE56/03	HF933423.1_CDS_3295	36.81	99.44	99.43
IP21447	CTFQ01000053.1_CDS_20	94.38	99.44	99.44
IP21447	CTFQ01000001.1_CDS_197	36.81	99.44	99.43
YE119/02	CTIM01000056.1_CDS_20	94.38	99.44	99.44

Table 5.3: BLASTP outputs showing the presence of *ail* and *ail* homologs in *Y. enterocolitica* strains, continued.

Strain name	Subject locus tag	Query sequence identity (%)	Query sequence coverage (%)	Subject sequence coverage (%)
YE119/02	CTIM01000003.1_CDS_98	36.81	99.44	99.43
1127	CPWQ01000113.1_CDS_1	94.38	99.44	99.44
1127	CPWQ01000005.1_CDS_93	94.38	99.44	99.44
1127	CPWQ01000003.1_CDS_96	36.81	99.44	99.43
2/C/53NMD7	CPWH01000103.1_CDS_2	94.38	99.44	99.44
2/C/53NMD7	CPWH01000074.1_CDS_7	36.81	99.44	99.43
YE11/03	CTJL01000066.1_CDS_2	94.38	99.44	99.44
YE11/03	CTJL01000001.1_CDS_96	36.81	99.44	99.43

From the BLASTP outputs, I found that every *Y. enterocolitica* strain had an *ail* homolog which was approximately 35% identical to functional *ail*. This suggests that the homolog is belonged to core gene family. Within highly pathogenic phylogroup, *Y. enterocolitica* ATCC 9610 (in red bold text) was the only strain which only had one *ail* homolog, suggesting the functional *ail* was lost. Such atypical observation was also present in non-pathogenic phylogroup, where I found that *Y. enterocolitica* YE53/30444 (in blue bold text) was the only strain which had two *ail* homologs while the rest non-pathogenic strains only had one. One of the homologs was 36.81% identical to functional *ail*, but another one was 73.68% identical to functional *ail*. To get more detailed understanding, I had also performed TBLASTN search against *Y. enterocolitica* genomes. The TBLASTN outputs are tabulated in Table 5.4, after discarding hits which also present in BLASTP outputs (see Table 5.3) unless the hit was overlapping with another hit within the same genome.

Table 5.4: TBLASTN outputs showing where the functional *ail* of *Y. enterocolitica* 8081 was used as query to search genomes of *Y. enterocolitica*. Hits which also present in BLASTP output (see Table 5.3) were discarded unless the hit was overlapped with another hit within the same genome.

Strain	Subject accession	Identity (%)	Query start position	Query end position	Subject start position	Subject end position	E-value
YE53/30444	CQEI01000020.1	79.07	93	178	47486	47743	2.00E-64
YE53/30444	CQEI01000020.1	73.68	1	93	47205	47489	2.00E-64

From the TBLASTN outputs, I found that only *Y. enterocolitica* YE53/30444 was left after filtering. I also found that the two hits were overlapping each other. For instance, in the first row, the hit started at position 47,486 while the hit at second row ended at position 47,489; there was an overlap for four nucleotides. By referring to a previous study which proposed methodology to identity bacterial pseudogene (Lerat & Ochman, 2004), the overlapping region could be an evidence of pseudogenization of *ail* in YE53/30444. A more detailed explanation is illustrated in Figure 5.5.

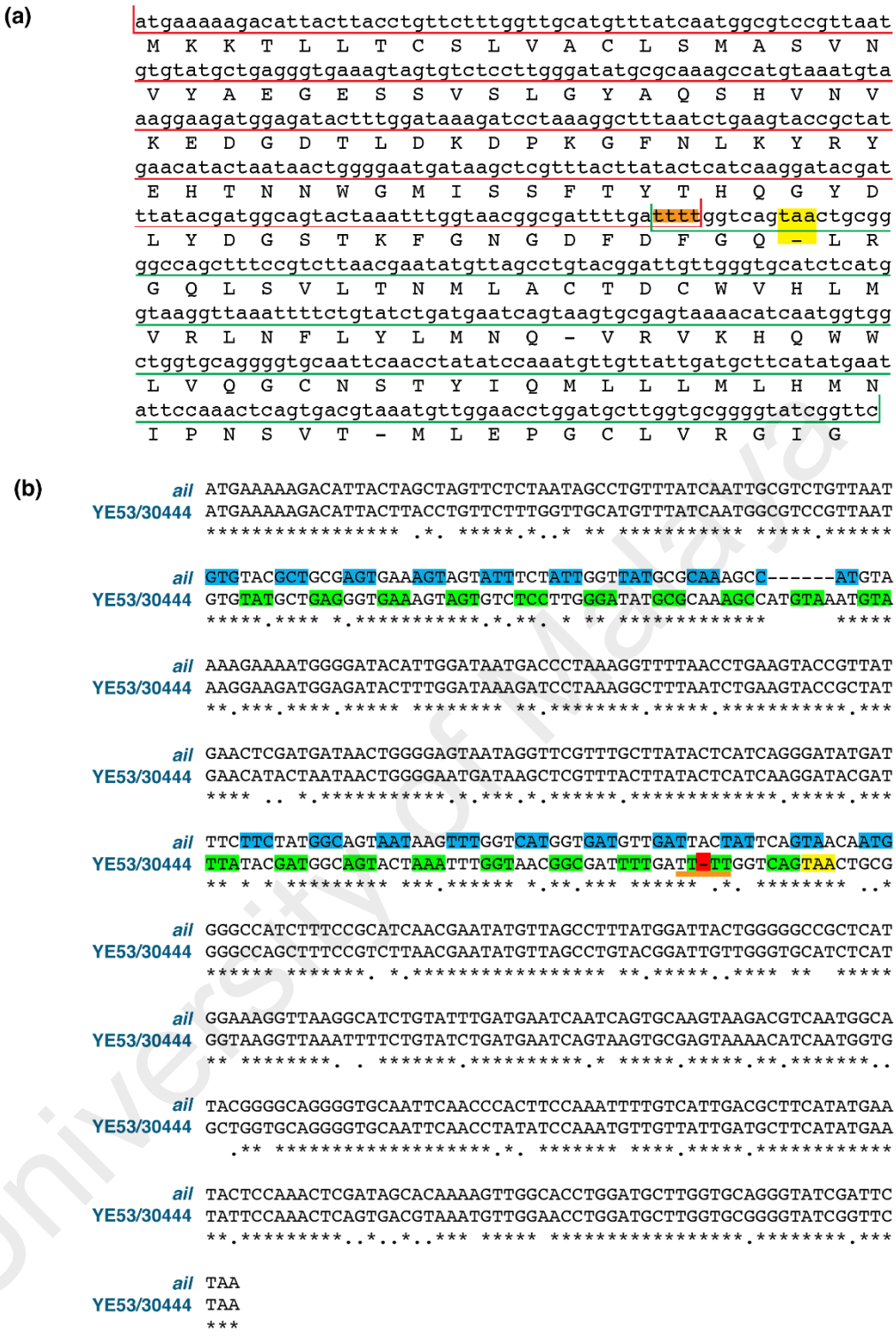

Figure 5.5: (a) TBLASTN mapped regions in the *Y. enterocolitica* YE53/30444 genomes were merged and translated into amino acids sequence. The two mapped regions were underlined by red and green colour, respectively. Overlapped region of the two hits is highlighted in orange while the stop codon adjacent to the region is highlighted in yellow. (b) TBLASTN mapped regions in the YE53/30444 genomes were merged and aligned with functional *ail* sequence of highly pathogenic *Y. enterocolitica* 8081. Codons adjacent to gap are highlighted in alternate blue-white and green-white colours. Premature stop codon is highlighted in yellow and putative frameshift mutation which adjacent to the stop codon is highlighted in red.

I found that there were several insertions/deletions in the alignments (Figure 5.5b). For instance, the insertions/deletions at the second row were triplets, indicating that they do not disrupt reading frames and are unlikely to affect the protein functions. However, there was a single gap (highlighted in red) at the fifth row of the alignments. In an uninterrupted reading frames, blue codon of *ail* should align with white codon of the subject (i.e. YE53/30444), and white codon of *ail* should align with green codon of the subject. I found that soon after the gap, blue codon was able to align partially with white codon, indicating that the gap was a single insertion/deletion and it likely to disrupt the reading frames. I also found a premature stop codon (highlighted in yellow) just after the gap. Taken all of the observations together, I suggest that the functional *ail* has been pseudogenized in *Y. enterocolitica* YE53/30444.

CHAPTER 6: RESULTS (PART III): YERSINIABASE

6.1 Overview and functionalities

YersiniaBase, which can be accessed at http://yersinia.um.edu.my, is a specialized platform designed to store genomic information of *Yersinia* strains and allow comparative analyses using online bioinformatics tools incorporated in the platform. All of the genomic features were stored into four different tables in MySQL relational database version 14.12 (http://www.mysql.com). Attributes of each table are tabulated in Table 6.1.

Table 6.1: Attributes of tables used to store genomic features of *Yersinia* strains in MySQL relational database.

Table name	Attributes
Species	Species name
Strain	Strain name Genome assembly status Genome size Number of contig Number of ORF Number of tRNA Number of rRNA Guanine-cytosine percentage
Feature	Type of ORF ORF start and stop positions Length of nucleotide sequence Length of amino acid sequence Sense (positive or negative strand) Function of ORF Subsystem Subcellular localization Hydrophobicity (pH) Molecular weight (Da)
Sequence	Nucleotide sequence Amino acid sequence

On entering the home page of YersiniaBase (Figure 6.1), visitors can view the news & conferences, blogs & information and the most recent published papers which are related to *Yersinia*. The "Browse" menu allows the visitors to view the list of *Yersinia* species currently available in YersiniaBase, with each "View Strains" button leading the visitors to the "Browse Strains" page, displaying all available strains of that respective species. In the "Browse Strains" page a general description of that particular species is given along with a table listing the strains of that species. Each strain is linked to their corresponding taxonomic classification page in NCBI and also to their page in Genome Online Database (GOLD) (Bernal et al., 2001). Furthermore, by clicking on the "Details" icon, visitors can obtain more comprehensive information of that particular strain such as their source and time of isolation, which we retrieved from NCBI, along with the list of ORFs, their respective function, start and stop positions in a tabular fashion in the "Browse ORF" page. Apart from that, each ORF is linked to its corresponding UniProt page along with their ORF ID being linked to its corresponding page in NCBI. By clicking on the Contig ID of each ORF, the corresponding contig information available in NCBI can be accessed. The Details button of each ORF leads the visitor to the "ORF Detail" page displaying the detailed information of that ORF such their type, start and stop positions, lengths of nucleotide as well as amino acid sequences. It further provides information on functional classification, subsystem (if available), strand, subcellular localization, hydrophobicity (pH) and molecular weight (Da). The page also displays the amino acid and the nucleotide sequence of the ORF along with the Genome Browser. The "Genome Browser" menu links user to JBrowse in YersiniaBase. JBrowse allows users to view the position of each ORF at each genome graphically (Skinner et al., 2009).

Figure 6.1: Home page of YersiniaBase which can be accessed at http://yersinia.um.edu.my.

I also integrated in-house tools and as well as other tools into YersiniaBase to add functionalities to this *Yersinia* research platform. The "Tools" menu allows user to perform a BLAST search against the *Yersinia* strains curated in YersiniaBase, as well as an exclusive BLAST search against the VFDB (Chen et al., 2012; Chen et al., 2005; Yang et al., 2008). Besides the BLAST search, users can also perform pairwise alignment of

any two *Yersinia* genomes present in YersiniaBase of their choice by using the PGC, draw a heat map of the virulence genes profiles by using the PathoProT or construct phylogenetic tree by using YersiniaTree.

"Search" menu allows user to search the functional classification of a specific species and strain by providing keyword or ORF ID. Besides performing searching, user can also download the genome sequence, ORF annotation details in table format, ORF sequence, ribonucleic acid (RNA) and coding sequence (CDS) through "Download" menu. The overview of the functionalities of YersiniaBase is shown in Figure 6.2 .

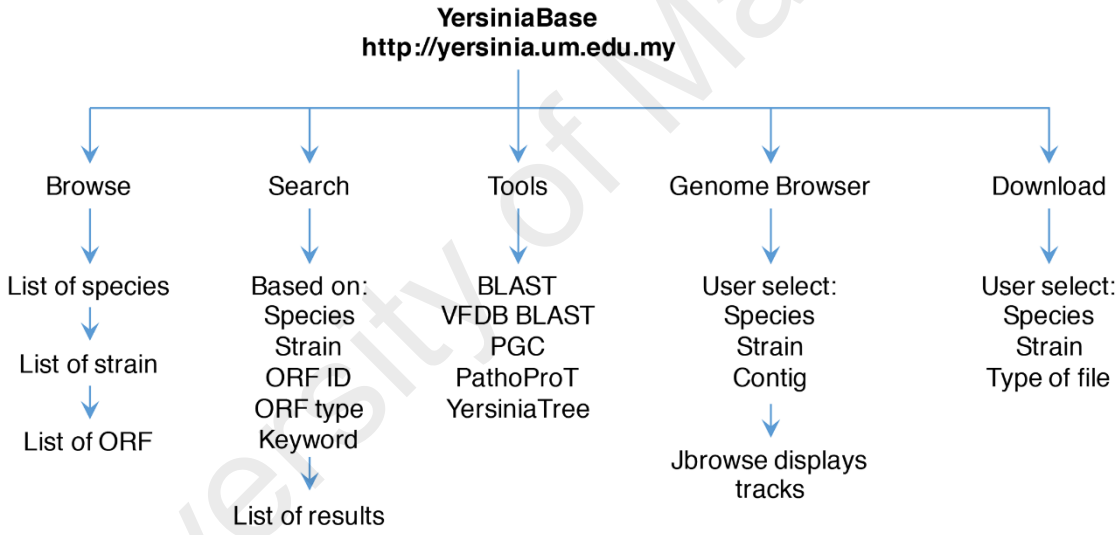

Figure 6.2: Overall functionalities of YersiniaBase.

6.2 Browsing genomic data in YersiniaBase

As all of the genomic data were linked to each other through relational relationships, YersiniaBase allows user to browse them level by level (Figure 6.3).

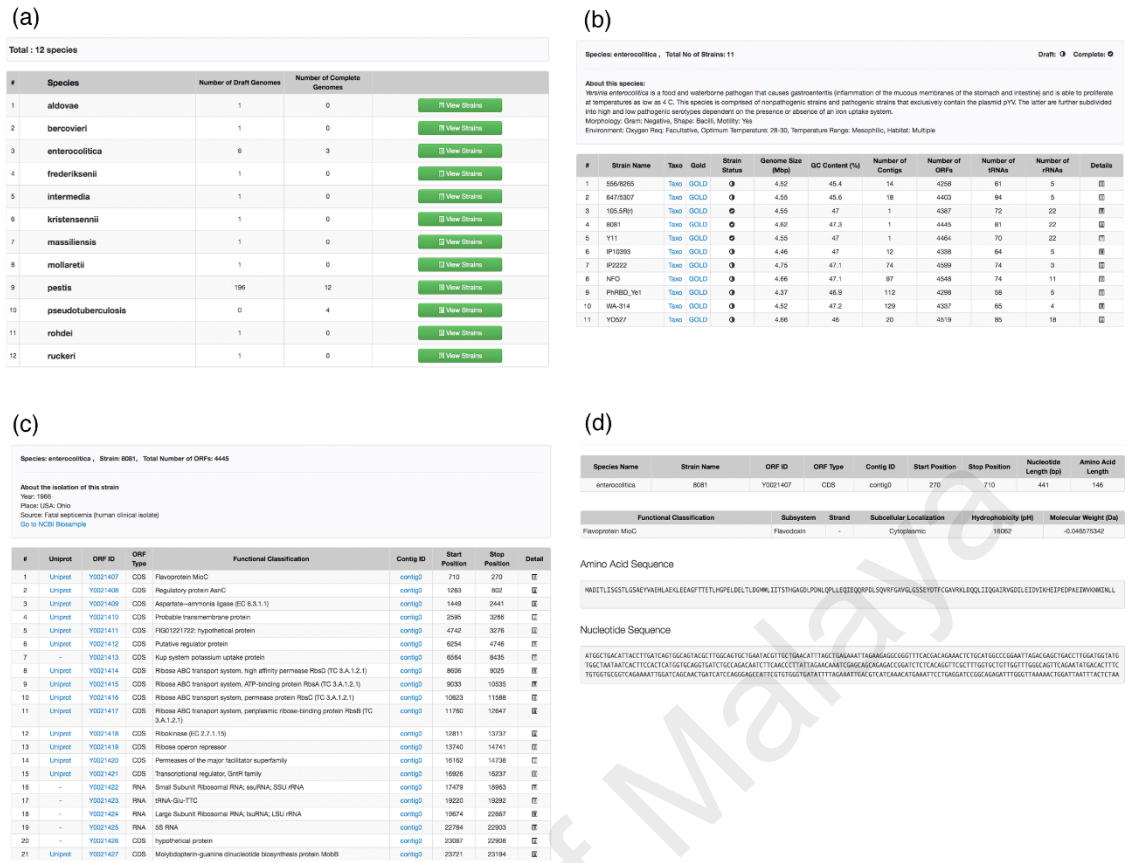

Figure 6.3: (a) Browsing list of species in YersiniaBase (b) Browsing list of strain of selected species (c) Browsing list of genes of selected strain (d) Browsing detailed information of a selected gene.

The browse page of YersiniaBase summarizes list of *Yersinia* species and their respective number of draft and complete genomes. The "View Strains" button on the right of each species leads the user to a list of strains of the chosen species. In the table, the information available to the user are strain status (draft genome or complete genome), genome size in mega base pair (Mbp), percentage of guanine-cytosine content (%), number of contigs, number of predicted CDS, number of predicted tRNA and number of predicted rRNA. On the right of each strain, the user can find a small icon, which provides a hyperlink enabling the user to find a list of predicted ORFs of the selected strain. From there, the user can see the ORF type (CDS or RNA), functional classification, contig, start position and stop position associated with each predicted ORF of the strain. The small icon on the right of each ORF brings the user to another page which allows them to see additional information of the predicted ORF besides that described above, which includes nucleotide

length (bp) and the sequence, predicted polypeptide length (amino acids) and the direction of transcription, subcellular localization, hydrophobicity (pH), molecular weight (Da) and SEED subsystem. The page is also equipped with JBrowse a fast and modern JavaScript-based genome browser which will enable the user to navigate genome annotations and visualize the location of the ORF of the selected *Yersinia* strain (Skinner et al., 2009). On the top of the page, the user can find a "Download" button to download annotation details, amino acid sequence and nucleotide sequence of the predicted ORF.

6.3 Real-time searching in YersiniaBase

As YersiniaBase stores more than one million genes and coding sequences related to *Yersinia*, it would impractical to search for the required information page by page as this will greatly slow down the progress. In order to counter this problem, I implemented a real-time search engine using AJAX in YersiniaBase. The real-time search engine was designed in such a way that the communications between web interface and MySQL database in asynchronous; refreshing of web page in unneeded to display the list of suggested functional classifications that match the entered keyword. The real-time search engine can be found in home page and is illustrated in Figure 6.4.

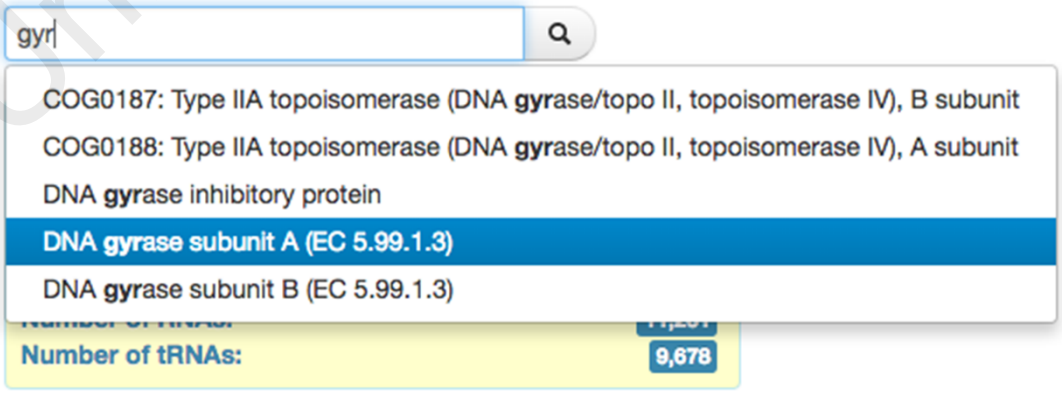

Figure 6.4: Real-time search engine in YersiniaBase which speeds up the process of searching for a specific gene.

In the real-time searching box, as soon as user enters one keyword, the real-time search engine will retrieve a list of functional classifications that contains the keyword entered by the user and display to the user seamlessly. For example, to search for gyrase related genes, by just typing "gyr", the system will list down all of the gene names which have "gyr". After users clicking on, for example, "DNA gyrase subunit A (EC 5.99.1.3)", they will be presented with a list of strains which have the gene of interest.

6.4 Pairwise Genome Comparison tool for genome wide comparison

To understand *Yersinia* genus, further study is required not only for the pathogenic strains but also non-pathogenic strains to gain a clear understanding of their biology. To have a clear idea of *Yersinia* genetics, an extended view of the gene pool and genomic information from a single *Yersinia* genome is unlikely to be sufficient. For detailed insights into the variations between different *Yersinia* strains at the genetic level, the evolutionary changes among *Yersinia* species, as well as the genes that give each strain its unique characteristics, especially the potential regions associated with pathogenicity, a comparative study of multiple *Yersinia* genomes is required. With that in mind, Pairwise Genome Comparison (PGC) tool is developed and incorporated into YersiniaBase.

PGC allows user to compare two *Yersinia* genomes of interest and displays the alignment in a circular layout. On entering the web interface of PGC in YersiniaBase, users can choose two *Yersinia* genomes of interest from the list for comparison. Alternatively, users can upload their *Yersinia* genome sequence for comparison with the *Yersinia* genomes available in YersiniaBase. There are three parameters which can be manually defined by users, including minimum percentage of identity, merge threshold (base pair) and link threshold (base pair). "Link" is the region where query genome maps to reference genome and is shown if the mapped region is higher than the value set in "link threshold". "Merge"

is the merging of two "links" which are separated by gap, into one wider "link". Two "links" are merged if the gap between them is shorter than the "merge threshold". An example is illustrated in Figure 6.5.

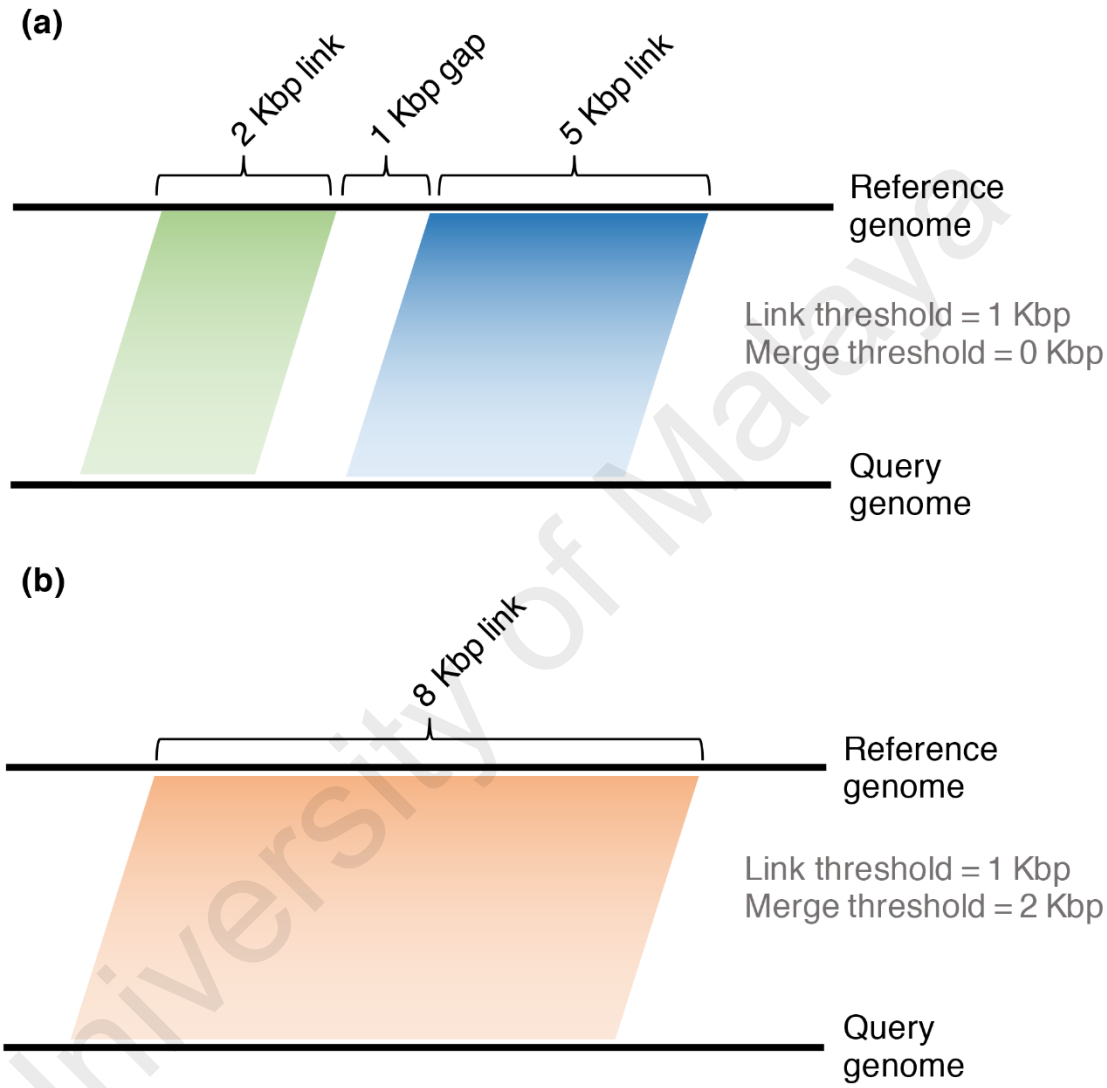

Figure 6.5: The effects of different parameters set in PGC tool. (a) Green and blue links are displayed as the mapped region, because the mapped region is higher than the link threshold, while the gap is present between green and blue link because the gap is wider than the value of merge threshold (0 Kbp in this case). (b) Since the gap (1 Kbp) is smaller than 2 Kbp (merge threshold in this case), the green and blue links beside the gap are merged into a wider link of 8Kbp (2 Kbp Green Link + 1 Kbp Gap + 5 Kbp Blue link).

In the PGC pipeline, two *Yersinia* sequences of interest are aligned using NUCmer, and the alignment results are parsed to Circos, which then generates circular layout to show the pairwise relationships between the two *Yersinia* sequences, with karyotypes and links encoding the position, size and orientation of the related genomic elements. (Delcher et al., 2002; Krzywinski et al., 2009). Perl scripts are used to automate the multi-step process of this pipeline. The results generated by PGC, which include NUCmer alignment and Circos diagram can be downloaded in PGC result page. Figure 6.6 illustrates the work flow of PGC tool, describe the integration of both MUMmer and Circos, generating the required input files for the PGC to function.

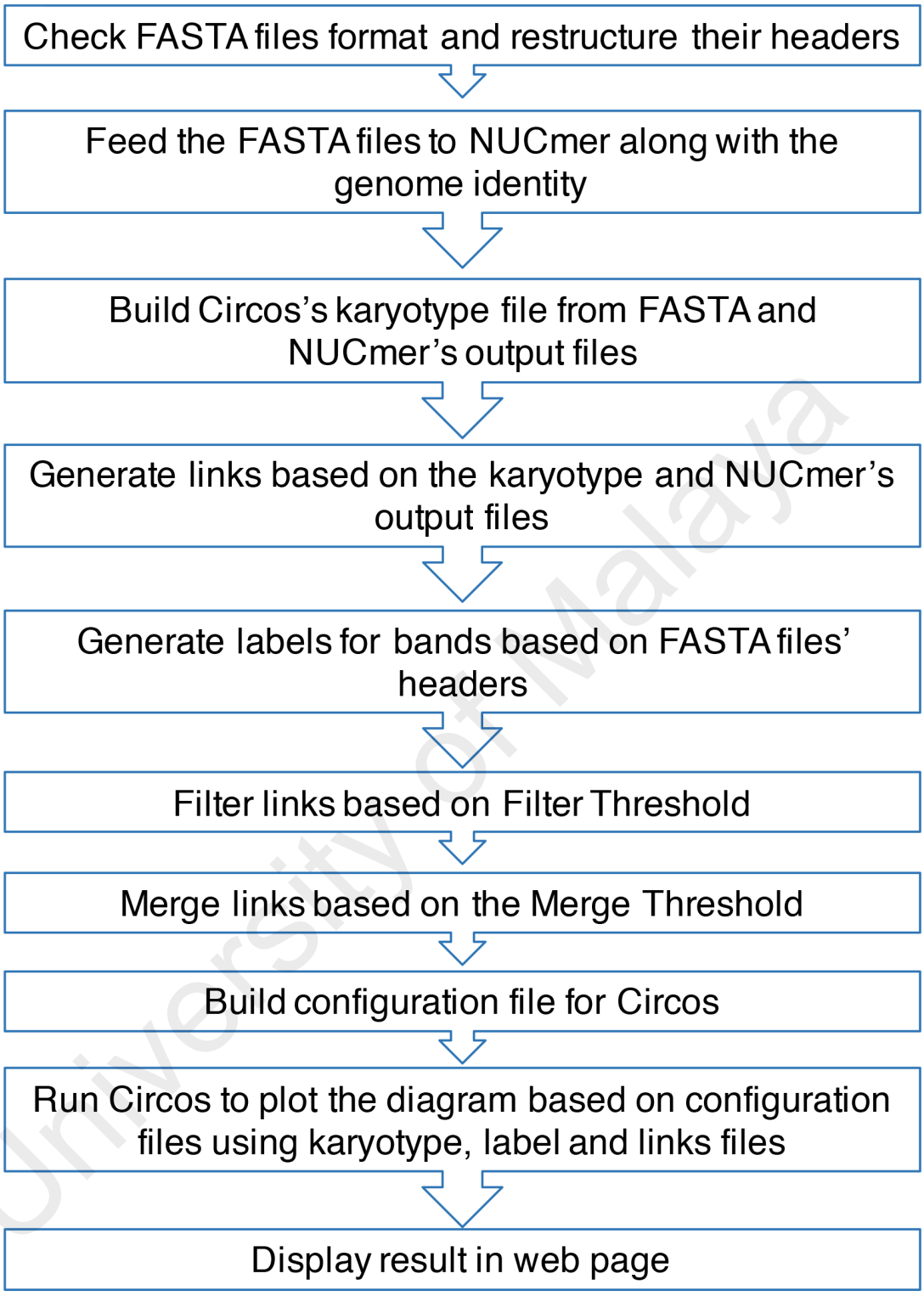

Figure 6.6: Description of processes taken in PGC pipeline after user submits the job to the server.

At the time of writing this thesis, a similar tool named Circoletto already exists, which aligns two genomes by using BLAST (Darzentas, 2010), however PGC aligns two genomes by using NUCmer package in MUMmer 3.0 (Delcher et al., 2002). In comparison, the latter is more favourable and more suitable for whole-genome comparison as NUCmer uses global alignment which is more suitable for large-scale and rapid pairwise alignment between two large genomes while Circoletto uses BLAST, a local alignment program (Delcher et al., 2002). PGC provides a user-friendly interface and requires no prior programming knowledge. PGC allows the user to adjust parameters such as minimum percent genome identity (%), merging of links/ribbons according to merge threshold and also the removal of links according to the user-defined link threshold through the provided online form. An example to show usage of PGC in *Yersinia* analysis is illustrated in Figure 6.7.

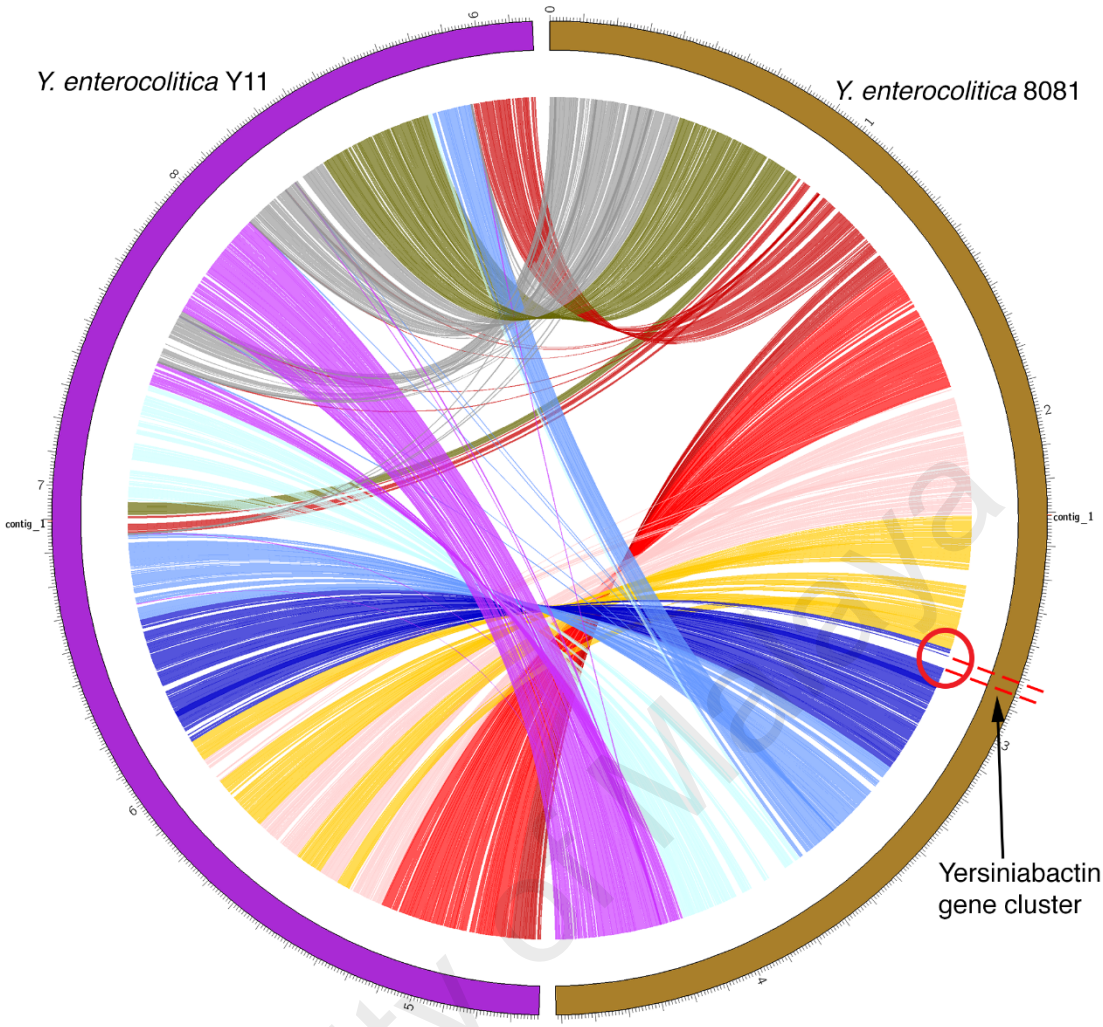

Figure 6.7: Pairwise Genome Comparison (PGC) tool aligned genomes between *Y. enterocolitica* 8081 and Y11, and showing region of yersiniabactin gene cluster in 8081 was not mapped by Y11.

The above example shows that PGC tool is useful in finding genomic differences between two *Yersinia* strains. The region of the operon for yersiniabactin in *Y. enterocolitica* 8081 is circled in red. From the figure generated by PGC, I found that *Y. enterocolitica* Y11 did not have region mapped to the region encoding yersiniabactin in 8081. It has been known that *Y. enterocolitica* Y11 is a low pathogenic strain and does not have yersiniabactin gene cluster (Pelludat et al., 1998).

6.5 Pathogenomics Profiling Tool for comparative virulence gene analysis

Pathogenesis is due to the presence of virulence genes in bacteria, which is responsible for causing disease in the host (Peterson, 1996). With the availability of genome sequences of different *Yersinia* species, it is essential to do comparative analyses of the virulence genes in *Yersinia* pathogen genomes to identify new potential virulence markers and to provide new insights into pathogenicity of this genus. In order to identify potential virulence genes in the *Yersinia* strains and to facilitate comparative analyses between different *Yersinia* strains, I have developed Pathogenomics Profiling Tool (PathoProT), a unique comparative virulence gene analysis tool implemented in YersiniaBase to help users to study pathogenicity of different species of *Yersinia*. It was designed using Perl and R scripts, where Perl handles the initial processes while R is used to generate the heat map. Users are allowed to select a list of *Yersinia* strains for comparative analysis and set the cut-off for sequence identity and completeness through online form in the PathoProT main page. PathoProT predicts virulence genes based on sequence homology of all protein sequences of user selected strains against VFDB using BLASTP (Altschul et al., 1990; Chen et al., 2012; Chen et al., 2005; Yang et al., 2008).

In the PathoProT pipeline, BLASTP search is performed with the default parameters of 50% sequence identity and 50% sequence completeness to identify homologs of these known virulence genes in the *Yersinia* genomes present in YersiniaBase. However, users can change these default parameters for the BLASTP search depending on their desired levels of stringency. In-house developed Perl script filters the results generated from BLASTP search against VFDB based on user-defined cut-off values for sequence identity and completeness to identify the virulence genes and selects only the user desired strains. The filtered results are then used to tabulate data matrix which is strains versus virulence genes. This is followed by executing R scripts to read data matrix and generate heat map to visualize virulence gene profiles. Figure 6.8 illustrates the flow chart of PathoProT, briefly describing the pipeline which integrates both Perl script and R script, and the processes before generating the output file.

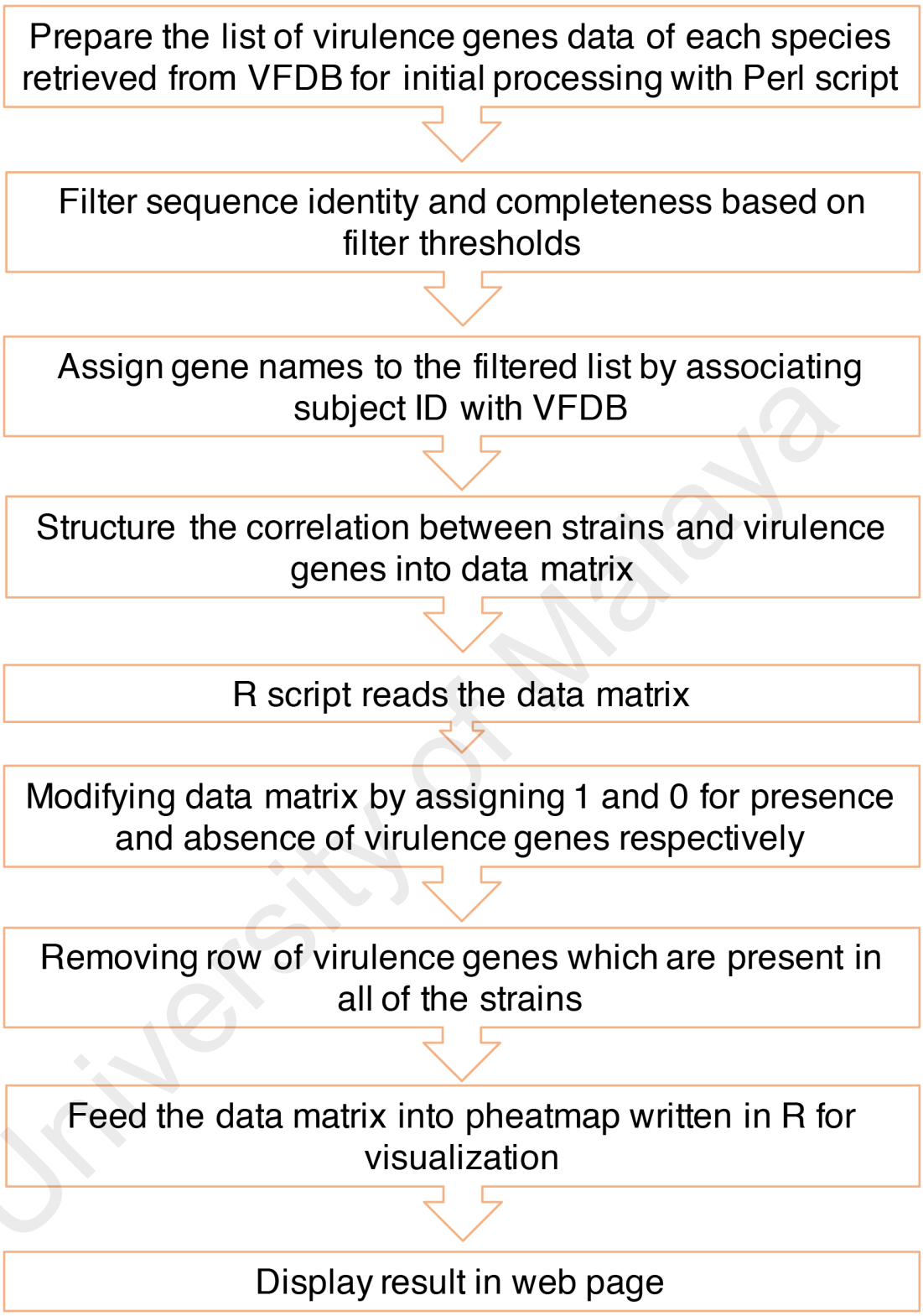

Figure 6.8: Description of processes taken in PathoProT pipeline after user submits the job to the server.

The use of heat map enables users to view the result in graphical representation, allowing comparative analyses of virulence genes among different *Yersinia* strains. An example to show usage of PathoProT is illustrated in Figure 6.9.

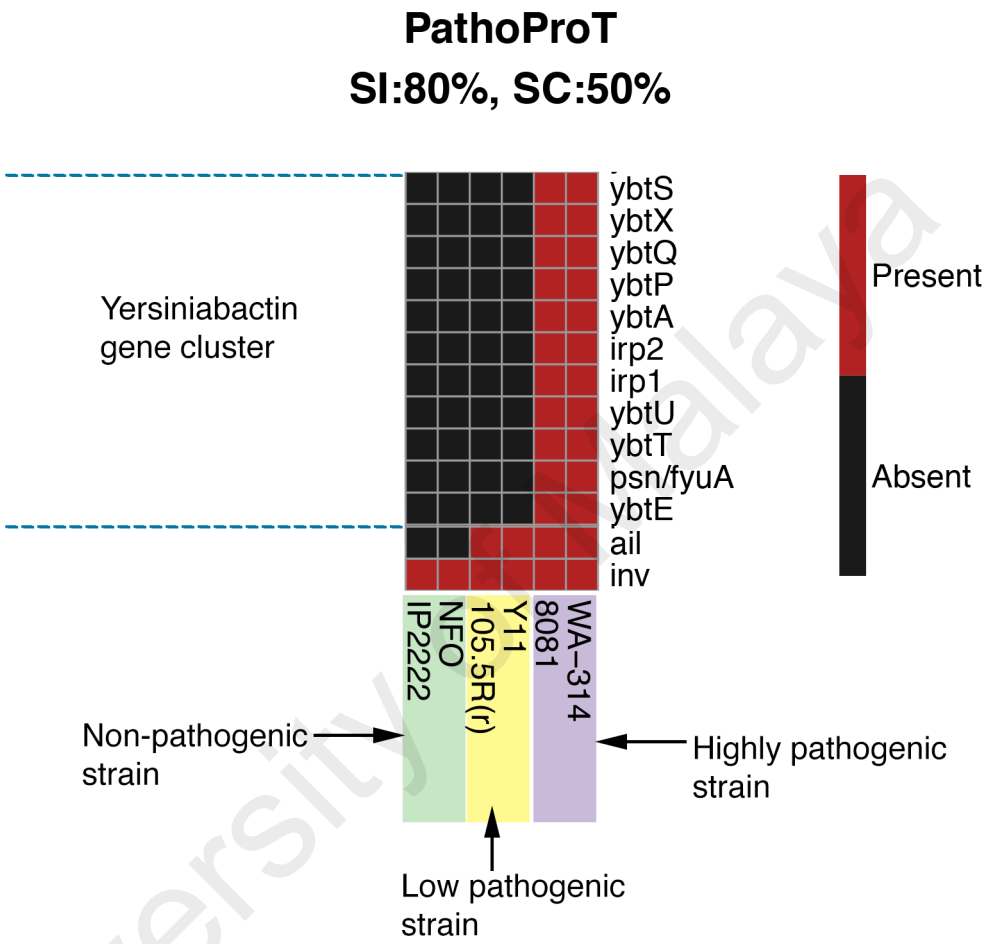

Figure 6.9: Example heat map generated by PathoProT showing presence and absence of virulence genes in six *Y. enterocolitica* strains. Yersiniabactin gene cluster was only present in highly pathogenic strain, *ail* was present in both highly pathogenic and low pathogenic strain while *inv* was present in all strains.

Figure 6.9 shows how PathoProtT can be used in comparative virulence gene analysis. Besides having pYV virulence plasmid, highly pathogenic *Y. enterocolitica* strains carry yersiniabactin genes (Pelludat et al., 1998). From the heat map, I found that yersiniabactin gene cluster was only present in highly pathogenic *Y. enterocolitica* strains, showing consistency with previous study (Pelludat et al., 1998). I also found that *ail* was present in both low pathogenic and highly pathogenic strains, while *inv* was present in all strains,

showing consistency in my results described above (see CHAPTER 5:RESULTS (PART II): THE SUBSPECIES OF *YERSINIA ENTEROCOLITICA*). Such graphical representation makes it easier and quicker for user to identify certain virulence genes in the selected strains.

6.6 YersiniaTree to construct *Yersinia* phylogenetic tree

YersiniaTree is an automated pipeline written in Perl and enables users to generate phylogenetic tree of *Yersinia* strains based on their housekeeping genes and 16S rRNA. YersiniaTree primarily requires two inputs from user: gene marker used to construct the phylogenetic tree and list of genomes in YersiniaBase which to be included in the tree. The automated pipeline also offers an optional feature where the users can input their nucleotide sequence in FASTA format along with the sequences which are retrieved from the database. Currently, YersiniaTree allows users to choose from one out of five gene markers, which are 16S rRNA and four housekeeping genes, including *gyr*B, *hsp*60, *rpo*B and *sod*A for the construction of phylogenetic tree. Previous studies have shown that phylogenetic trees based on these four housekeeping genes are more consistent with the biochemical profiles of *Yersinia* species (Merhej et al., 2008a; Stenkova et al., 2012). After user provides all of the necessary inputs, the front end PHP executes the backend Perl pipeline. The first step of the automated pipeline is to select target gene's nucleotide sequence of genomes chosen by user from FASTA file where complete listings of sequences are stored, into a temporary FASTA file. The Perl script then executes MAFFT, which is used to perform multiple sequence alignment across nucleotide sequences (Katoh & Standley, 2013). Next, the output file form MAFFT is sent to FastTree to construct phylogenetic tree in newick format, followed by visualizing of the tree using Newick Utilities (Junier & Zdobnov, 2010; Price et al., 2010). Finally, Perl script sends out the final image to PHP to display the phylogenetic tree in web browser.

6.7 Sequence-based searches

Besides new bioinformatics tools, I have also integrated BLAST and VFDB-BLAST into YersiniaBase. This allows users to perform similarity search of their query sequences against *Yersinia* genome sequences, gene sequences and virulence genes using BLAST. This specialized dataset for *Yersinia* enables faster searching against *Yersinia* sequences compared to NCBI NT or NR database when users are trying to find the closest *Yersinia* strains to their own sequences.

CHAPTER 7: DISCUSSION

7.1 Evolution of human pathogenic *Yersinia* species

The most recent and promising evolutionary study which elucidates the evolution of human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis* was performed by Reuter et al. (2014), which hypothesized that ecological speciation caused *Yersinia* species to evolve in parallel. However, prokaryotic evolution can be affected by many factors such as ecological specialization, gene gain-and-loss, gene duplication and lateral gene transfer (Jensen, 2001; Lassalle et al., 2015; Ochman et al., 2000; Ravenhall et al., 2015). In this study, I have successfully performed a series of analyses including phylogenetic tree analysis, gene gain-and-loss analysis, recombination testing, CRISPR analysis and the virulence gene homolog analysis, which have given better and more comprehensive insights into the evolution of human pathogenic *Yersinia* species. I found that the evolution of human pathogenic *Yersinia* was a multifactorial process, instead of solely resulting from ecological specialization as proposed by Reuter et al. (2014).

Firstly, an accurate and robust phylogenetic tree is important to infer the phylogenetic relationships between the *Yersinia* species. Hence, I have chosen to construct a supermatrix tree based on non-recombinant super-sequence which is free of recombination. This is because recombinant genes can overwrite the history of vertically-transferred orthologs, and distort the topology of phylogenetic trees, making them unreliable (Fraser et al., 2007). In my approach, I used super-sequences, which concatenated alignments of multiple genes to reconstruct the phylogenetic trees. This approach can provide more phylogenetic signals to construct robust phylogenetic tree compared to the single gene approach (e.g. 16S rRNA) (de Queiroz & Gatesy, 2007). Using the supermatrix tree approach, I showed that *Y. enterocolitica* is distantly related

to *Y. pseudotuberculosis*-*Y. pestis*, and they likely evolved from different non-pathogenic populations. Moreover, as the *Yersinia* supermatrix tree was rooted in this study, it could provide more information compared to the previous studies, and is useful to trace the acquired and lost genes in ancestors before the emergence of human pathogenic *Yersinia* (Csuros, 2010).

Based on the gene gain-and-loss analysis, LCAHPY, which was the last common ancestor shared by human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*, seemed to have an environmental origin which might have adapted to live in the human gastrointestinal tract. For instance, the presence of *iol*ABCDEG (myo-inositol degradation genes) and *pga*ABCD (poly-beta-1,6-N-acetyl-D-glucosamine synthesis and transport genes) could allow it to survive in soil and vegetables, whereas the acquisition of the *ure*ABCEFGD (urease genes) could allow it to survive inside the human gastrointestinal tract if the host had consumed vegetables contaminated by LCAHPY. The adaptation to human gastrointestinal tracts could be an important milestone to give rise to new *Yersinia* lineages because the gastrointestinal tracts might provide a wide range of metabolic compounds and niches that can be readily exploited and invaded by the LCAHPY (Rohmer et al., 2011). For instance, if certain subpopulations of LCAHPY had acquired new metabolic capabilities through lateral gene transfer, they might have selective advantage in a particular niche compared to the other subpopulations. This further yielded nascent populations (or lineages) in new niches (Pal et al., 2005; Rohmer et al., 2011; Wiedenbeck & Cohan, 2011). My analyses clearly showed that both phylogroup-P and phylogroup-E species diverged from LCAHPY and had acquired new putative metabolism genes which likely utilize different nutrients in the human body.

Besides that, I found that phylogroup-E species had acquired new genes to utilize tetrathionate, 1-2-propanediol and hydrogen which are present in gastrointestinal tract. Previous studies showed that these genes allow *Salmonella* to outcompete other enteric bacteria and increase their fitness (Bobik et al., 1999; Maier et al., 2004; Price-Carter et al., 2001). This suggests that the phylogroup-E species especially the enteropathogenic *Y. enterocolitica* might have been successful in invading new niches inside human gastrointestinal tract compared to their ancestor and phylogroup-P. Thus, the phylogroup-E species could be viewed as a successful gastrointestinal colonizer that might have been transformed from an environmental species. On the other hand, the phylogroup-P species seemed to have expanded their ecological niche to macrophages, a view supported by the acquisition of putative genes such as *ter* and *rip* loci, and the loss of two *bcs* loci involved in cellulose biosynthesis. Such gained and lost genes were believed to have contributed to increased virulence inside the macrophages (Ponnusamy & Clinkenbeard, 2015; Ponnusamy et al., 2011; Pontes et al., 2015; Sasikaran et al., 2014). With the acquisition of these key genes, the phylogroup-P species might have acquired the capability to occupy the macrophage, a different body location compared to its predecessor (LCAHPY) and the phylogroup-E species which adapted well to the intestinal tracts (as suggested by my analyses). In such case, the adaptation to different body parts or locations could be a better and more efficient strategy to divide nutrients between the phylogroup-E and phylogroup-P. Therefore, I propose that the acquisition of different sets of metabolism genes by the phylogroup-P and phylogroup-E species is very important because of the following reasons: (1) it would ensure that one phylogroup decreases its ability to invade the niches of the other phylogroup and to compete for the same resources, thus allowing the phylogroups to evolve independently of each other (Lassalle et al., 2015); and (2) it could make one population to be less susceptible to the selection of another population, allowing beneficial mutations to be accumulated within its own population (Cohan, 2001).

Although both phylogroup-P and phylogroup-E seemed to have adapted to their respective niches and might have formed stable ecotypes, there is a possibility that recombination and lateral gene transfer can occur between both phylogroups. This is because gene loci that have neutral effects and do not deteriorate fitness of bacteria in their respective niches are known to be susceptible to recombination and maintained in the recipient genome (Lassalle et al., 2015). If the interspecies recombination in *Yersinia* is extensive, the process may remove the nucleotide variations resulted from mutations, and prevent genomic-wide divergence. When such scenario happens, one would find *Yersinia* to be a fuzzy species, whereby given a *Yersinia* species A, it would contain sequences from another *Yersinia* species B or C (Corander et al., 2012). This would probably further halt speciation and divergence of the three *Yersinia* phylogroups. Fortunately, in the estimation of rate of recombination to mutation, I found that mutation played a more dominant role over recombination to introduce genomic variations in *Yersinia*. On the other hand, to ensure a clear separation between phylogroup-P and phylogroup-E, newly acquired metabolism genes must be maintained within the respective phylogroup and not transferred to another phylogroup (for example, by lateral gene transfer) (Pal et al., 2005). If the unique metabolism genes from phylogroup-P are transferred laterally to phylogroup-E and vice versa, then one would find no distinct ecotype or niche between them as the two phylogroups would have adapted to each other's niches (Wiedenbeck & Cohan, 2011). My analyses showed that the gene content based phylogenetic tree exhibited highly similar phyletic patterns to the supermatrix tree, suggesting that lateral gene transfer might not be extensive between the two phylogroups and they maintain distinguishable gene content from each other (Snel et al., 1999). In short, I found that the rate of gene flow between phylogroup-P and phylogroup-E was likely low and that mutation likely played a major role in causing elevated nucleotide divergence in these *Yersinia* phylogroups. Therefore, sexual mating between these

phylogroups would be reduced as divergent sequences form the barrier to the process (Majewski et al., 2000).

7.2 Non-parallel evolution of human pathogenic *Yersinia*

Up to this point, I have explained how *Yersinia* phylogroups could evolve independently to each other by adapting to different niches. However, the ecological speciation process neither fully explains nor justifies the transformation of these ancestral species into present-day pathogenic species. Therefore, I hypothesize that there may be a series of events that have led to the emergence of human pathogenic *Yersinia* species, which could not just be explained by the independent evolution.

Contradicting to a recent study which hypothesized that human pathogenic *Yersinia* had evolved in parallel (Reuter et al., 2014), I found that the virulence *ail* gene was likely acquired through gene duplication and lateral gene transfer but not in parallel. Analyses on the *ail* homologs present in *Yersinia* species showed that the phylogroup-P pathogenic species had multiple copies of *ail* homolog, which might be duplicated in the genome of Ancestor_Yps (the last common ancestor of all *Y. pseudotuberculosis*-*Y. pestis* strains) after the divergence from the human non-pathogenic *Y. similis*. After the gene duplication, there could be redundant copies of the same gene which perform the same physiological role in Ancestor_Yps, rendering one (or some) of the duplicated genes to have weaker purifying selection and experienced multiple mutations (Kondrashov et al., 2002). As these paralogs were homologous to present-day *ail*, one could assume that they were also outer membrane proteins which might be able to interact with mammalian cell receptors. Thus, beneficial mutations that took place in one of the paralogs might increase the efficiency to bind and interact with host cell, while the rest of the genes could still perform the same physiological role as before. As a result, neofunctionalization of paralog might

have happened and facilitated the emergence of *ail*. These arguments show consistency with a previous study which found that most of the duplicated genes were membrane and secretion genes (Kondrashov et al., 2002).

One interesting biological question is how did pathogenic *Y. enterocolitica* acquire the *ail* gene? My data showed that there were one functional *ail* and one *ail* homolog in the *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*. If the *Y. enterocolitica* had acquired *ail* independently from the *Y. pseudotuberculosis*-*Y. pestis*, its *ail* gene should have higher sequence identity to its own *ail* homolog compared to the homologs from other *Yersinia* species. However, I found that the *ail* gene of *Y. enterocolitica* showed highest sequence identity to the *ail* and the *ail* homologs of *Y. pseudotuberculosis* (as a reference of phylogroup-P) instead of to its own homologs. This suggests that the phylogroup-P species might be the donor of the *ail* virulence gene to *Y. enterocolitica*, for example, through lateral gene transfer event.

Besides the *ail* gene, I also studied the evolution of another virulence gene, *inv*. My analyses suggest that *inv* homologs were likely inherited from LCAHPY. The *inv* homolog of LCAHPY was probably not related to virulence because there were no reports that descendants of LCAHPY (except human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*) which inherited the LCAHPY's *inv* homolog, are able to use their *inv* homologs to invade human cell lining. Thus, transformation of non-virulence-related *inv* homolog into functional *inv* that is involved in pathogenesis might have resulted from adaptive mutations leading to functional change in the ancestral *inv* homolog. This hypothesis is made on the basis that a recent report has found the relationships between adaptive mutations and virulence in *Salmonella* (Chattopadhyay et al., 2012). Recombination analysis showed that probability of recombination in *inv* gene family was high, suggesting functional *inv* might have been transferred laterally between

Y. enterocolitica and *Y. pseudotuberculosis*. The possibility of lateral transfer of *inv* between the two species is also supported by the same N-terminal present in their *inv* but absent in other *Yersinia* species, as indicated by BLASTP output. However, unlike genes homologous to functional *ail* which are present in multiple copies in human pathogenic *Yersinia* species, gene homologous to *inv* is only present in single copy in each of these species. Due to the absence of second copy of *inv* homolog, I could not perform pairwise sequence comparison (which is similar to *ail* homologs) between functional *inv* genes and *inv* homologs to determine if functional *inv* of *Y. enterocolitica* is closer to its own *inv* homolog than to the *inv* homolog of *Y. pseudotuberculosis*. Hence, it is still unknown which human pathogenic *Yersinia* species is the donor of *inv*.

The next question is what factors have caused the acquisition of the pYV virulence plasmid in only human pathogenic *Yersinia* species but not in the other species? I found that the loss of CRISPR-Cas system could be one of the critical factors for the acquisition of the pYV virulence plasmid in the human pathogenic *Yersinia* species. For instance, in the phylogroup-E, the human non-pathogenic *Y. frederiksenii* and *Y. kristensenii* have spacers that can recognize pYE854 plasmid, which is able to mobilize pYV plasmid (Hammerl et al., 2008). The immunity to pYV is achieved probably through fragmentation of pYE854-pYV cointegrate. I also found that the human pathogenic *Y. enterocolitica*, which also belong to phylogroup-E, had lost the CRISPR-Cas system. This might allow *Y. enterocolitica* to acquire pYV plasmid and transform into a human pathogen.

On the other hand, in the phylogroup-P, my data showed that the spacers in the pathogenic *Y. pseudotuberculosis*-*Y. pestis* could only recognize the pYV harboured by *Y. enterocolitica* but not their own pYV plasmid. This shows the high specificity of CRISPR-Cas system in *Yersinia* because if the spacers of *Y. pseudotuberculosis*-*Y. pestis*

can target a common region found in every pYV plasmid, their pYV plasmids will be fragmented by their own CRISPR-Cas system (Horvath & Barrangou, 2010; Marraffini, 2013). I cannot rule out the possibility that independent mutations might have caused variations in the pYV sequences harboured by the *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*, allowing the CRISPR-Cas system in the latter to have good precision when targeting foreign DNA materials.

An interesting question is why some human non-pathogenic *Yersinia* species such as *Y. ruckeri* has not acquired the pYV plasmid but yet it does not have the CRISPR-Cas system? Answers to this question may lie in the properties of pYV, selection that acts on pYV and the cost to bear a plasmid (Baltrus, 2013; San Millan et al., 2014). There are two prerequisites for the expression of Ysc-Yop T3SS: 37°C and direct contact to the host cells (Cornelis et al., 1998). For instance, even if *Y. ruckeri*, which mainly associates with rainbow trout and lives in aquatic environment that does not reach 37°C (Romalde & Toranzo, 1993), has accidentally acquired pYV plasmid, the plasmid would not increase its fitness or virulence to infect fish. This is because with the lower temperature in the aquatic niche of the rainbow trout, the pYV plasmid in *Y. ruckeri* (if it existed) could not be activated or become functional. Thus the pYV plasmid would likely be negatively selected for and eventually become lost.

Besides that, my analyses found that some human non-pathogenic *Yersinia* species such as *Y. aldovae*, *Y. aleksiciae*, *Y. intermedia* and *Y. rohdei* also do not have pYV plasmid. It is known that these species do not have functional adhesins such as *inv*, *ail* and *psa*, which commonly associate with pathogenesis, the host cell attachment and the Yop delivery (Mikula et al., 2012). For instance, although some of these non-pathogenic species have the *inv* homologs, the homologs are likely non-functional due to the lack of proper N-terminal which is required for proper localization of Inv (Leong et al., 1990).

Therefore, these *Yersinia* species, even when they have the pYV plasmid, might not be able to cross epithelial cell lining to reach lymph nodes and cause disease like pathogenic *Y. enterocolitica*. Taken all together, I believe that the loss of CRISPR-Cas system is crucial to the acquisition of the pYV virulence plasmid by *Yersinia* species. However, due to high cost for bacteria to bear plasmid (Baltrus, 2013), there are also two important factors to determine if the pYV is favoured by selection and would be maintained in bacterial cells: presence of 37°C in the environment and presence of virulence genes to assist in the pathogenesis.

Although *Y. enterocolitica* is distantly related to *Y. pseudotuberculosis*-*Y. pestis* (about 81% identical to each other's chromosomes), I found that their pYV plasmids showed high ANI values (> 96%), suggesting that their pYV plasmids might have a single origin. Since *Y. enterocolitica* does not share the same direct ancestor with *Y. pseudotuberculosis*-*Y. pestis*, it is unlikely that their plasmids originated from the common ancestor. As the pYV plasmid is only found in human pathogenic *Yersinia* species (Cornelis, 2002a), it is likely to originate from either *Y. enterocolitica* or *Y. pseudotuberculosis*-*Y. pestis*. The question is which, *Y. enterocolitica* or *Y. pseudotuberculosis*-*Y. pestis*, was the first to acquire the pYV virulence plasmid and became the donor of pYV to the other human pathogenic *Yersinia* species?

My analyses suggest that the spacers which could recognize the pYV plasmid might have originated from non-*Yersinia* species and probably existed before the acquisition of the pYV plasmid. This is because the top BLAST hits were not the pYV plasmid but plasmids from other genus such as *Escherichia* and *Clostridium*, suggesting that the donors of these pYV-recognizing spacers likely originated from non-*Yersinia.* Nevertheless, these spacers could provide immunity towards the pYV plasmid of *Yersinia* species due to their

sequence similarity (Ravenhall et al., 2015). Thus, several hypothetical scenarios could be made to infer which *Yersinia* species has first acquired the pYV plasmid:

- I hypothesize that *Y. pseudotuberculosis*-*Y. pestis* might be the first to acquire the pYV plasmid and then transfer it to *Y. similis*. However, due to the presence of spacers in *Y. similis* which could recognize the pYV plasmid, the latter was likely to be fragmented. Thus, *Y. similis* remains non-pathogenic and has no pYV plasmid.

- *Y. pseudotuberculosis*-*Y. pestis* might also successfully transfer their pYV plasmid to *Y. enterocolitica* after the latter has lost the CRISPR-Cas system. As time passed, the pYV plasmid harboured by *Y. enterocolitica* has slight sequence variations in certain regions due to mutations.

- *Y. enterocolitica* might attempt to transfer its own pYV plasmid to the other *Yersinia* species. As the spacers in *Y. similis* are able to recognize conserved regions in the pYV plasmid of *Y. enterocolitica*, its CRISPR-Cas system could still fragment the plasmid. While in the *Y. pseudotuberculosis*-*Y. pestis*, their CRISPR-Cas system have spacers which could recognize the mutated region in pYV of *Y. enterocolitica*, they could also fragment the pYV of *Y. enterocolitica.* Besides, it would be costly to bear redundant copies of pYV plasmid (San Millan et al., 2014).

- If *Y. enterocolitica* is the first species to acquire the pYV plasmid and it attempts to transfer pYV to *Y. pseudotuberculosis*-*Y. pestis*, the plasmid would be likely fragmented by the latter due to the presence of the CRISPR-Cas system. Thus, it

would be impossible for *Y. pseudotuberculosis*-*Y. pestis* to possess pYV in present-day.

7.3 Evolutionary model of human pathogenic *Yersinia* species

Taken all of my arguments together, I hypothesize that the evolution of human pathogenic *Yersinia* is a multifactorial process: ecological specialization, gene duplication, loss of CRISPR-Cas system and lateral gene transfer have contributed to *Yersinia* evolution. Based on a series of comparative analyses and evolutionary studies, I am able to propose a more complete and robust evolutionary history, which is shown in Figure 7.1, to elucidate the emergence of human pathogenic *Yersinia* species compared to previous studies (Reuter et al., 2014; Wren, 2003).

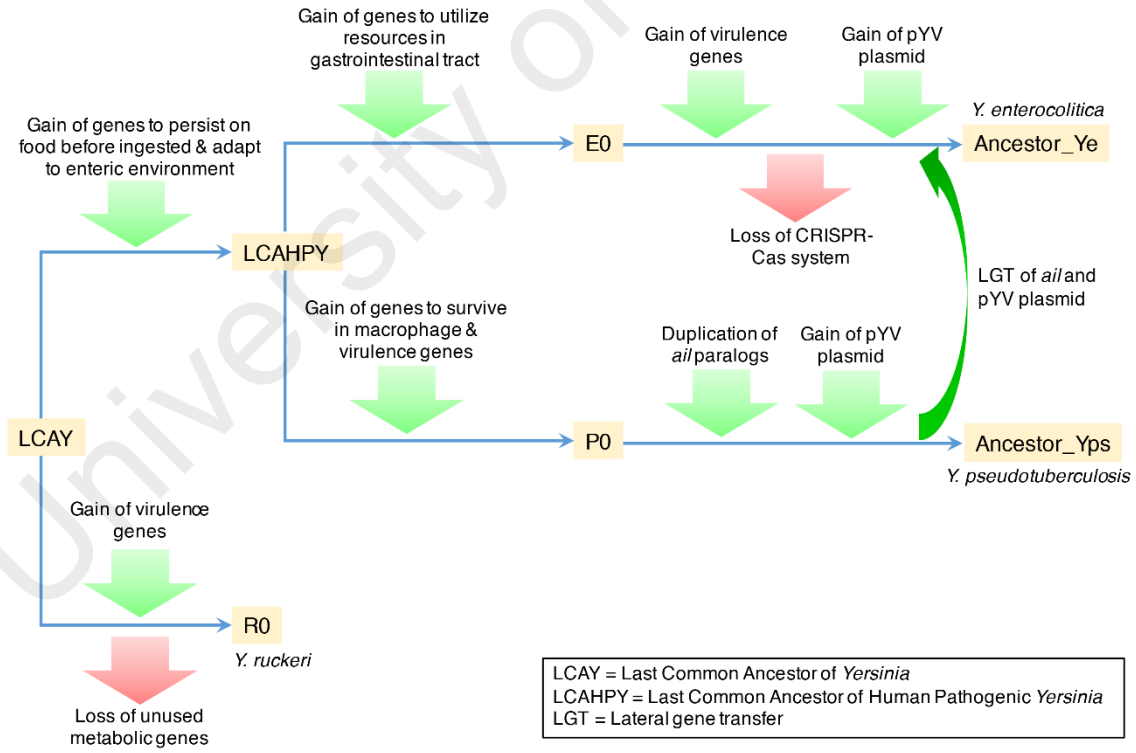

Figure 7.1: Key evolutionary events that might have occurred in *Yersinia* which led to the emergence of human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*.

In chorological order, the possible key evolutionary events which might have taken place in the past are:

(1) The emergence of LCAHPY, which might have an environmental origin and became the last common ancestor shared by the human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*, through the acquisition of genes to survive inside human gastrointestinal tract.

(2) Diversification of LCAHPY took place through ecological specialization. Each subpopulation might have gained different genes and developed strategies to metabolize different nutrients available in different niches and body locations.

(3) Emergence of phylogroup-P and phylogroup-E from LCAHPY with reduced recombination and lateral gene transfer. Both phylogroups might have formed stable ecotypes in their respective ecological niches.

(4) Gene duplication took place in *Y. pseudotuberculosis*-*Y. pestis,* allowing them to acquire the *ail* genes and transform into pathogens. The *Y. pseudotuberculosis*-*Y. pestis* might have also gained the pYV virulence plasmid.

(5) The lateral gene transfer of the *ail* virulence gene from *Y. pseudotuberculosis*-*Y. pestis* to *Y. enterocolitica*.

(6) *Y. enterocolitica* lost the CRISPR-Cas system and immunity to the pYV plasmid. *Y. pseudotuberculosis*-*Y. pestis* might be the donor of the pYV plasmid to *Y. enterocolitica* through lateral gene transfer.

(7) Positive selection and maintenance of pYV plasmid inside *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*.

7.4 Subspeciation in *Yersinia enterocolitica*

Y. enterocolitica can cause human infections (Bottone, 1997), and the roles of the virulence genes and the pYV plasmid in conferring pathogenicity have been well-studied (Batzilla et al., 2011a; Cornelis, 2002a; Mikula et al., 2012; Pelludat et al., 1998). However, the evolution of *Y. enterocolitica* especially its subspecies is still largely unknown. The most up-to-date study regarding the subspecies of *Y. enterocolitica* is the one conducted by Howard and colleagues, in which they proposed that there were three subspecies in *Y. enterocolitica* (Howard et al., 2006). Similar to the *Yersinia* genus, I postulate that the evolution and subspeciation of *Y. enterocolitica* was likely due to various factors. Hence, I have performed more detailed analyses in order to elucidate the evolution of this species. My findings suggest that the evolution of *Y. enterocolitica* was accompanied by ecological specialization, pseudogenization and gene gain-and-loss.

First of all, I have constructed a robust and accurate phylogenetic tree to infer the phylogenetic relationships between *Y. enterocolitica* strains using supermatrix tree, instead of 16S rRNA phylogenetic tree mainly because of two reasons:

- The non-recombinant super-sequence which I have used to infer the supermatrix tree had 1,138,594 nucleotides which was much longer than 16S rRNA which has about 1532 nucleotides. This indicates that the non-recombinant super-sequences will provide more information and more phyletic signals which can generate a more accurate phylogenetic tree (de Queiroz & Gatesy, 2007).

- Although the 16S rRNA was previously used to propose two subspecies of *Y. enterocolitica*, which were *Y. enterocolitica* subsp. *palearctica* and *Y. enterocolitica* subsp. *enterocolitica* (Neubauer et al., 2000), many studies have shown that the 16S rRNA lacks phylogenetic power to infer relationships between

Yersinia and the other bacteria (Dewhirst et al., 2005; Merhej et al., 2008b). Therefore, the classification of two subspecies might not be accurate and requires reassessment using a more robust phylogenetic tree.

Using a rooted supermatrix tree, I have successfully inferred the evolutionary relationships between *Y. enterocolitica* strains and the order of speciations. For instance, all of the strains used in this study were clearly demarcated into three distinct phylogroups, probably representing three subspecies of *Y. enterocolitica*. Contrary to a previous study (Howard et al., 2006), my approach indicates that the high pathogenic phylogroup was not the direct descendant of the most recent ancestor of all *Y. enterocolitica* strains (I have designated as Ancestor_Ye in this study). Instead, the non-pathogenic phylogroup was the direct descendent of Ancestor_Ye. The possible cause to account for this difference is that previous study did not use outgroup to root the phylogenetic tree and thus the order of subspeciation could not be inferred correctly. I found that there were two possible important subspeciation events that had taken place in the past in *Y. enterocolitica* species:

- In the first subspeciation, the population of Ancestor_Ye had diverged into two populations: pathogenic population and nonpathogenic population.

- In the second subspeciation, the pathogenic population further evolved into highly pathogenic population and low pathogenic population.

From my analyses described in the first result section (see CHAPTER 4:RESULTS (PART 1): THE HUMAN PATHOGENIC *YERSINIA* SPECIES) where I have studied the emergence of human pathogenic *Y. enterocolitica* and *Y. pseudotuberculosis*-*Y. pestis*, I found that the ancestral *Y. enterocolitica* was already equipped with many putative metabolism genes, probably to exploit different nutrients in human gastrointestinal tract.

Thus, new niches might continue to provide the driving force to further trigger ecological specialization or subspeciation. This might explain the acquisition of *dsd* and *fuc* loci by Ancestor_Nonpathogenic or the *aat* locus by the Ancestor_LowPathogenic. These newly acquired metabolism genes were not acquired by phylogroup-P, where *Y. pseudotuberculosis*-*Y. pestis* belonged to, as shown in my analyses. This supports the view that the ecological specialization which took place during the early divergence of LCAHPY might have allowed the phylogroup-E and phylogroup-P species to acquire distinct metabolism genes separately, but also decreased their ability to invade each other's niches.

Although ANI can be used to identify a species, it is not proved to be suitable to identity subspecies (Konstantinidis & Tiedje, 2005; Richter & Rossello-Mora, 2009). Nevertheless, one could assume that subspeciation is a process which is similar to speciation, whereby gene flow will be reduced so that there is less cohesive force between subpopulations and each subpopulation becomes more specialized to its own niches (Fraser et al., 2007; Lassalle et al., 2015). As time passes, these subpopulations could diverge from each other and become subspecies of *Y. enterocolitica*. One of the approaches is to create phylogenetic network to visualize if there is any conflicting signal which spans across different subpopulations as recombinations tend to form cohesive forces between subpopulations (Huson & Bryant, 2006). The *Y. enterocolitica* phylogenetic network that I constructed in this study has exhibited the expected topology whereby most of the reticulations were found within phylogroup, rather than between phylogroups. This suggests that recombination tends to take place among *Y. enterocolitica* strains belonging to the same phylogroup and thus there is higher cohesive force between them compared to the strains from different phylogroups. Furthermore, the estimation of the rate of recombination to mutation has indicated that mutations are dominant over the recombination, suggesting the frequency of recombination will be

reduced as the genomes between phylogroups become more diverged (Majewski et al., 2000). I also found that the gene content-based phylogenetic tree showed the existence of three phylogroups, which are similar to the topology in the supermatrix tree generated in this study. This suggests that the strains belonging to the same phylogroup may harbour genes unique to their own, and that lateral gene transfer across phylogroups may not be extensive (Snel et al., 1999). Due to distinguishable gene content and less lateral gene transfer, strains from different phylogroups likely exhibit different physiological traits, which could explain the heterogeneous properties of *Y. enterocolitica* (Bottone, 1997; Thomson et al., 2006). In short, I believe that subspeciation might have taken place in *Y. enterocolitica,* and lower rate of gene flow likely caused the process to be irreversible.

Up to this point, I have argued that the speciation in *Yersinia* and subspeciation in *Y. enterocolitica* followed the similar evolutionary styles, whereby subpopulations adapted themselves to new niches, resulting in a decreasing gene flow. I have also argued that the human pathogenic *Y. pseudotuberculosis*-*Y. pestis* might be the donor of pYV virulence plasmid and *ail* virulence gene to *Y. enterocolitica*. However, several questions arose:

- There were four important ancestors, which were Ancestor_Ye, Ancestor_Pathogenic, Ancestor_HighPathogenic, Ancestor_LowPathogenic and Ancestor_Nonpathogenic. Of these four, which of them was likely to become the recipient of the pYV plasmid and *ail*?

- Which of the four ancestors was the progenitor of present-day human pathogenic *Y. enterocolitica* strains?

By analyzing the *ail* homologs in the *Y. enterocolitica* strains, I found that the functional *ail* might be once present in *Y. enterocolitica* YE53/30444, but it was pseudogenized due to frameshift mutations. In bacterial genome, a disrupted gene is generally not functional and it is eliminated from the genome (Ochman & Davalos, 2006). Thus, the most parsimonious explanation for not being able to find functional *ail* in non-pathogenic strains is that the non-pathogenic *Y. enterocolitica* might have been very effective in removing pseudogenized gene(s) from its genome. Hence, it is possible that (1) the *ail* gene was gained by Ancestor_Ye and inherited by the Ancestor_Pathogenic and Ancestor_Apathogenic, and (2) the *ail* gene was pseudogenized in the non-pathogenic strains before it was removed in a relatively short time.

From the gene gain-and-loss analysis, my data clearly showed that the Ancestor_Ye has lost the CRISPR-Cas system and is likely the first ancestor to acquire pYV plasmid. The next question is how did the non-pathogenic phylogroup lose the pYV plasmid? The emergence of non-pathogenic *Y. enterocolitica* ATCC 9610 from the highly pathogenic phylogroup might provide some hints. My analyses show that the *Y. enterocolitica* ATCC 9610 has evolved from a highly pathogenic phylogroup which have *inv*, *myf*, *ail*, *yts*1-T2SS, *ysa*-T3SS and pYV plasmid (Bottone, 1997; Cornelis, 2002a). Among these virulence genes, only the *ail* and pYV were lost in *Y. enterocolitica* ATCC 9610. In a recent study, the *ail* gene has been reported to be the only chromosomal locus that is involved in the delivery of Yop proteins (encoded by pYV plasmid) in *Y. enterocolitica* (Mikula et al., 2012). This could reflect the importance of the *ail* and pYV plasmid and also a possible reason to explain why they were lost together in ATCC 9610. I suggest that a similar scenario could be applied in the non-pathogenic phylogroup, whereby after pseudogenization of *ail*, the pYV plasmid might be gradually lost due to lack of *ail* protein to assist in the Yop delivery. Taken all together, I speculate that Ancestor_Ye is likely a pathogen and progenitor of the pathogenic *Y. enterocolitica*.

The impact of ecological specialization on subspeciation of *Y. enterocolitica* should not be neglected, but geographic isolation seems necessary to be evoked in order to explain the divergence between the highly pathogenic and low pathogenic phylogroups because their classification could be associated with geographical areas (Howard et al., 2006). Previous study proposed that the movement of continents in the distant past has caused the split between the low pathogenic and highly pathogenic *Y. enterocolitica* strains (Howard et al., 2006). However, the hypothesis was drawn without inferring to the properties of Ancestor_Pathogenic, which was the most recent ancestor of highly pathogenic and low pathogenic phylogroups. As discussed above, Ancestor_Ye had gained various virulence determinants such as *ail*, *inv* and pYV plasmid. My gene gain-and-loss analysis showed that none of these genes were lost in the Ancestor_Pathogenic. This suggests that the Ancestor_Pathogenic was a pathogenic species due to the presence of important virulence genes, allowing it to either infect hosts, or survive in its natural reservoir, just like present-day pathogenic *Y. enterocolitica* (Schaake et al., 2014; Valentin-Weigand et al., 2014). Following my hypothesis, there could be two possible factors that caused the geographical isolation between sub-populations of Ancestor_Pathogenic and triggered their divergence:

- Geographical factor: rifted continents between North America and Eurasia that separated these subpopulations.

- Vector-mediating factor: Migration of infected hosts or natural reservoir from Eurasia to North America or vice versa.

7.5 Evolutionary model of subspeciation in *Yersinia enterocolitica*

Taken all together, I hypothesize that the subspeciation has occurred in *Y. enterocolitica* and the process is affected by multiple factors such as the ecological specialization, the loss of CRISPR-Cas system, the *ail* pseudogenization and geographical isolation. Based on a series of comparative analyses and evolutionary studies, I would like to propose a more complete and robust evolutionary history to illustrate the evolution and subspeciation of *Y. enterocolitica*, compared to previous studies (Howard et al., 2006; Neubauer et al., 2000; Thomson et al., 2006), which is shown in Figure 7.2.

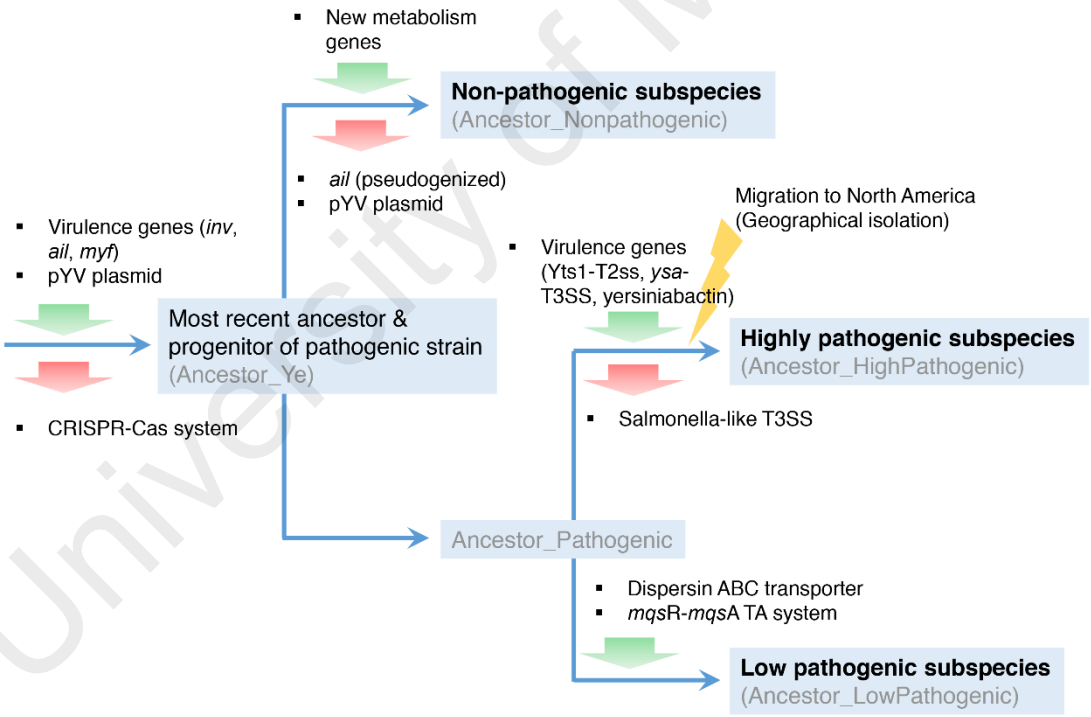

Figure 7.2: Key evolutionary events that likely took place in *Y. enterocolitica* and led to the emergence of non-pathogenic subspecies, low pathogenic subspecies and highly pathogenic subspecies.

Based on this hypothesis, the key evolutionary events might have taken place in the past in the following chorological order:

(1) Ancestor_Ye has already adapted to human gastrointestinal tract. At this point, it gained the *ail* virulence gene from the *Y. pseudotuberculosis*-*Y. pestis* (phylogroup-P species) through lateral gene transfer. It also lost CRISPR-Cas system and acquired pYV virulence plasmid.

(2) Subpopulations of Ancestor_Ye continued invading new niches and subspeciated into the pathogenic and non-pathogenic subspecies. Gene flow between these two subspecies decreased to prevent the population of one sub-species to invade the niches of the other sub-species.

(3) The emergence of Ancestor_Nonpathogenic and non-pathogenic subspecies was probably accompanied by pseudogenization of *ail* and loss of pYV plasmid, which then rendered the populations to become harmless to human.

(4) Some of the strains belonging to pathogenic subspecies might be brought into North America, either due to rifted continents or migration of host or reservoir.

(5) Pathogenic *Y. enterocolitica* strains which were isolated in North America continued to evolve by acquiring new virulence genes and become highly pathogenic subspecies. At the same time, another group of pathogenic strains in Eurasia evolved independently and became low pathogenic subspecies.

(6) With each of the subspecies evolved in parallel without invading each other's niches, each of them exhibits different physiological traits. *Y. enterocolitica* is now consisting of heterogeneous collection of strains.

7.6 YersiniaBase for *Yersinia* research community

With the advent of next generation sequencing technologies, it is important that biological data can be stored and retrieved easily for analyses. YersiniaBase aims to provide the first platform that stores genomic data and annotation details of *Yersinia* besides providing new in-house designed bioinformatics tools particularly for comparative analyses that can accelerate research in *Yersinia*. I have successfully demonstrated how these tools can be used in analyses including the identification of pathogenicity factors and genomic differences.

With the rapid advances in NGS technologies and significant price drop in sequencing, I expect more *Yersinia* strains will be sequenced and published in future. Thus, the data in YersiniaBase will be continued updating once these genome sequences and computing resources are available in public. I hope that YersiniaBase will provide a comprehensive resource and analysis platform for of the *Yersinia* research community in the future.

7.7 Biological significance and future direction

This comparative study has used a series of bioinformatics approaches to study *Yersinia* genomes from various perspectives to elucidate the emergence of human pathogenic *Yersinia* species and also the subspeciation of *Y. enterocolitica* in detailed, which is not reported previously. From the evolutionary perspective, my findings can provide better insights to *Yersinia* research community by elucidating how the evolution of *Yersinia* is affected by different factors and thus, not entirely in parallel. Of all the bioinformatics approaches employed, the gene gain-and-loss analysis might be helpful in providing useful information for microbiologists, who focus on microbiology and bacterial metabolism research. For instance, the adaptation to human gastrointestinal tract would be the early step that led to emergence of human pathogenic *Yersinia* species. Before

expressing any virulence traits or infecting host cells, these *Yersinia* pathogens would have to increase their fitness by adapting and metabolizing available nutrients found in the environment (Rohmer et al., 2011). This suggests that metabolism is the prerequisite for virulence in pathogenic *Yersinia*. Although there have been studies that describing relationships between metabolic capabilities and virulence in other bacteria genus (Connolly et al., 2016; Maier et al., 2004; Pontes et al., 2015; Wu et al., 2012), analyses on *Yersinia* pathogens are still favouring over characterization of virulence genes (Hammerl et al., 2008; Mikula et al., 2012; Rakin et al., 2012; Schaake et al., 2014; Valentin-Weigand et al., 2014). These have resulted in poor understanding on the most fundamental aspects of human pathogenic *Yersinia*, which are the metabolic pathways used by them to ensure survival. Thus, the gene gain-and-loss analysis can become a reference for future experiments. For example, microbiologists could study the effect of losing *rip*ABC (itaconate catabolism genes) on *Y. pseudotuberculosis*-*Y. pestis* to see if there is any decrease in virulence level as I found that their emergence was accompanied with the gain of *rip* locus, which might be beneficial to their survival and virulence.

The analyses on virulence gene homologs bring concerns to the non-pathogenic subspecies of *Y. enterocolitica*, which consists of biotype 1A and is generally considered to be non-pathogenic due to absence of pYV virulence plasmid (Bottone, 1997). Recently, *Y. enterocolitica* biotype 1A strains have been isolated from a few clinical cases and it is debatable whether they are pathogenic to human (Fredriksson-Ahomaa et al., 2012; Kanaujia et al., 2015; Singhal et al., 2016; Stephan et al., 2013). In this study, I found the *inv* homologs in the nonpathogenic subspecies. These *inv* homologs were found to have at least 84% identity and 100% query coverage to functional *inv*, suggesting both of the genes might be functionally similar. A previous study has demonstrated that some of the biotype 1A strains were able to invade epithelial cells, suggesting the presence of *inv* homolog might be one of the key mechanisms. In addition to that, gene gain-and-loss

analysis showed that several new metabolism genes were gained by the non-pathogenic subspecies, such as *dsd* and *fuc* loci. These metabolism genes have been associated with virulence of other human pathogens (Connolly et al., 2015; Stahl et al., 2011). This hypothesizes that biotype 1A strains might be able to proliferate inside our gastrointestinal tract if our food intake contains metabolic compounds that are readily metabolized by them. Thus, biotype 1A strain should not be regarded as totally non-pathogenic. I suggest that phenotypic studies be conducted to determine the correlation between food and the metabolic activities of biotype 1A strains, and their potential pathogenicity to human. Until clear correlations are obtained, this supposedly non-pathogenic strain should be handled with caution.

Last but not least, I also wish to highlight that further studies should be performed to characterize *Y. similis*, which is generally thought to be non-pathogenic due to the lack of pYV virulence plasmid (Sprague & Neubauer, 2014). My gene gain-and-loss analysis showed that many metabolism genes related to persistence in macrophages, which were present in human pathogenic *Y. pseudotuberculosis*-*Y. pestis*, were also found in *Y. similis*. Further analyses also showed that the homologs of most typical *Yersinia* virulence genes such as *pil* locus, *psa* locus, *ail* and *inv*, were all also present in the *Y. similis* genome. It has been demonstrated that *Yersinia* species that have no pYV virulence plasmid can still be invasive due to the presence of virulence genes in their chromosomes (Fukushima et al., 1991; Grant et al., 1998; Lian et al., 1987). Although the human clinical case caused by *Y. similis* has not been reported so far, the presence of virulence-associated metabolism genes and homologs of classical *Yersinia* chromosomal virulence genes may make it prudent to study this species and monitor its potential pathogenicity in the future.

CHAPTER 8: CONCLUSION

In this study, I have successfully performed evolutionary studies and comparative analyses on human pathogenic *Yersinia* species and subspecies of *Y. enterocolitica* strains through phylogenetic tree analysis, gene gain-and-loss analysis, recombination analysis, analysis of virulence genes homologs and the CRISPR-Cas system analysis. Based on the new findings and knowledge obtained from these analyses, I have proposed that the emergence of human pathogenic *Yersinia* and subspecies of *Y. enterocolitica* are a multifactorial process rather than just evolved in parallel as previously thought Their evolutions are affected by ecological specialization, gene duplication, lateral gene transfer and gene pseudogenization as elucidating in more complete models proposed in this study. As from the biological perspective, this study also suggests that the acquisition of new metabolism genes is important in ecological specialization for *Yersinia*. In addition, to accelerate the research in *Yersinia*, I have also developed YersiniaBase, a specialized genomic resource and comparative analysis platform for research community. Overall, this study has provided better and more comprehensive insights into the evolution of human pathogenic *Yersinia* and subspecies of *Y. enterocolitica*.

REFERENCES

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., & Carniel, E. (1999). Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. *Proc Natl Acad Sci U S A, 96*(24), 14043-14048.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., & Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics, 9*, 75. doi:10.1186/1471-2164-9-75

Baltrus, D. A. (2013). Exploring the costs of horizontal gene transfer. *Trends Ecol Evol, 28*(8), 489-495. doi:10.1016/j.tree.2013.04.002

Batzilla, J., Antonenka, U., Hoper, D., Heesemann, J., & Rakin, A. (2011a). Yersinia enterocolitica palearctica serobiotype O:3/4--a successful group of emerging zoonotic pathogens. *BMC Genomics, 12*, 348. doi:10.1186/1471-2164-12-348

Batzilla, J., Hoper, D., Antonenka, U., Heesemann, J., & Rakin, A. (2011b). Complete genome sequence of Yersinia enterocolitica subsp. palearctica serogroup O:3. *J Bacteriol, 193*(8), 2067. doi:10.1128/JB.01484-10

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2015). GenBank. *Nucleic Acids Res, 43*(Database issue), D30-35. doi:10.1093/nar/gku1216

Bernal, A., Ear, U., & Kyrpides, N. (2001). Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res, 29*(1), 126-127.

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., & Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics, 8*, 209. doi:10.1186/1471-2105-8-209

Bobik, T. A., Havemann, G. D., Busch, R. J., Williams, D. S., & Aldrich, H. C. (1999). The propanediol utilization (pdu) operon of Salmonella enterica serovar

Typhimurium LT2 includes genes necessary for formation of polyhedral organelles involved in coenzyme B(12)-dependent 1, 2-propanediol degradation. *J Bacteriol, 181*(19), 5967-5975.

Bottone, E. J. (1997). Yersinia enterocolitica: the charisma continues. *Clin Microbiol Rev, 10*(2), 257-276.

Brocker, M., Schaffer, S., Mack, C., & Bott, M. (2009). Citrate utilization by Corynebacterium glutamicum is controlled by the CitAB two-component system through positive regulation of the citrate transport genes citH and tctCBA. *J Bacteriol, 191*(12), 3869-3880. doi:10.1128/JB.00113-09

Brown, B. L., Grigoriu, S., Kim, Y., Arruda, J. M., Davenport, A., Wood, T. K., Peti, W., & Page, R. (2009). Three dimensional structure of the MqsR:MqsA complex: a novel TA pair comprised of a toxin homologous to RelE and an antitoxin with unique properties. *PLoS Pathog, 5*(12), e1000706. doi:10.1371/journal.ppat.1000706

Bruen, T. C., Philippe, H., & Bryant, D. (2006). A simple and robust statistical test for detecting the presence of recombination. *Genetics, 172*(4), 2665-2681. doi:10.1534/genetics.105.048975

Cadillo-Quiroz, H., Didelot, X., Held, N. L., Herrera, A., Darling, A., Reno, M. L., Krause, D. J., & Whitaker, R. J. (2012). Patterns of gene flow define species of thermophilic Archaea. *PLoS Biol, 10*(2), e1001265. doi:10.1371/journal.pbio.1001265

Cao, J., Woodhall, M. R., Alvarez, J., Cartron, M. L., & Andrews, S. C. (2007). EfeUOB (YcdNOB) is a tripartite, acid-induced and CpxAR-regulated, low-pH Fe2+ transporter that is cryptic in Escherichia coli K-12 but functional in E. coli O157:H7. *Mol Microbiol, 65*(4), 857-875. doi:10.1111/j.1365-2958.2007.05802.x

Carniel, E. (2001). The Yersinia high-pathogenicity island: an iron-uptake island. *Microbes Infect, 3*(7), 561-569.

Carniel, E., Guilvout, I., & Prentice, M. (1996). Characterization of a large chromosomal "high-pathogenicity island" in biotype 1B Yersinia enterocolitica. *J Bacteriol, 178*(23), 6743-6751.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol, 17*(4), 540-552.

Caulfield, A. J., & Lathem, W. W. (2012). Substrates of the plasminogen activator protease of Yersinia pestis. *Adv Exp Med Biol, 954*, 253-260. doi:10.1007/978-1-4614-3561-7_32

Chain, P. S., Carniel, E., Larimer, F. W., Lamerdin, J., Stoutland, P. O., Regala, W. M., Georgescu, A. M., Vergez, L. M., Land, M. L., Motin, V. L., Brubaker, R. R., Fowler, J., Hinnebusch, J., Marceau, M., Medigue, C., Simonet, M., Chenal-Francisque, V., Souza, B., Dacheux, D., Elliott, J. M., Derbise, A., Hauser, L. J., & Garcia, E. (2004). Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. *Proc Natl Acad Sci U S A, 101*(38), 13826-13831. doi:10.1073/pnas.0404012101

Chattopadhyay, S., Paul, S., Kisiela, D. I., Linardopoulou, E. V., & Sokurenko, E. V. (2012). Convergent molecular evolution of genomic cores in Salmonella enterica and Escherichia coli. *J Bacteriol, 194*(18), 5002-5011. doi:10.1128/JB.00552-12

Chen, L., Xiong, Z., Sun, L., Yang, J., & Jin, Q. (2012). VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res, 40*(Database issue), D641-645. doi:10.1093/nar/gkr989

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., & Jin, Q. (2005). VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res, 33*(Database issue), D325-328. doi:10.1093/nar/gki008

Chen, P. E., Cook, C., Stewart, A. C., Nagarajan, N., Sommer, D. D., Pop, M., Thomason, B., Thomason, M. P., Lentz, S., Nolan, N., Sozhamannan, S., Sulakvelidze, A., Mateczun, A., Du, L., Zwick, M. E., & Read, T. D. (2010). Genomic characterization of the Yersinia genus. *Genome Biol, 11*(1), R1. doi:10.1186/gb-2010-11-1-r1

Choi, E., Groisman, E. A., & Shin, D. (2009). Activated by different signals, the PhoP/PhoQ two-component system differentially regulates metal uptake. *J Bacteriol, 191*(23), 7174-7181. doi:10.1128/JB.00958-09

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Res, 44*(D1), D67-72. doi:10.1093/nar/gkv1276

Cohan, F. M. (2001). Bacterial species and speciation. *Syst Biol, 50*(4), 513-524.

Cohan, F. M. (2002). What are bacterial species? *Annu Rev Microbiol, 56*, 457-487. doi:10.1146/annurev.micro.56.012302.160634

Collyn, F., Lety, M. A., Nair, S., Escuyer, V., Ben Younes, A., Simonet, M., & Marceau, M. (2002). Yersinia pseudotuberculosis harbors a type IV pilus gene cluster that contributes to pathogenicity. *Infect Immun, 70*(11), 6196-6205.

Connolly, J. P., Gabrielsen, M., Goldstone, R. J., Grinter, R., Wang, D., Cogdell, R. J., Walker, D., Smith, D. G., & Roe, A. J. (2016). A Highly Conserved Bacterial D-Serine Uptake System Links Host Metabolism and Virulence. *PLoS Pathog, 12*(1), e1005359. doi:10.1371/journal.ppat.1005359

Connolly, J. P., Goldstone, R. J., Burgess, K., Cogdell, R. J., Beatson, S. A., Vollmer, W., Smith, D. G., & Roe, A. J. (2015). The host metabolite D-serine contributes to bacterial niche specificity through gene selection. *ISME J, 9*(4), 1039-1051. doi:10.1038/ismej.2014.242

Corander, J., Connor, T. R., O'Dwyer, C. A., Kroll, J. S., & Hanage, W. P. (2012). Population structure in the Neisseria, and the biological significance of fuzzy species. *J R Soc Interface, 9*(71), 1208-1215. doi:10.1098/rsif.2011.0601

Cornelis, G. R. (2002a). The Yersinia Ysc-Yop 'type III' weaponry. *Nat Rev Mol Cell Biol, 3*(10), 742-752. doi:10.1038/nrm932

Cornelis, G. R. (2002b). The Yersinia Ysc-Yop virulence apparatus. *Int J Med Microbiol, 291*(6-7), 455-462.

Cornelis, G. R., Boland, A., Boyd, A. P., Geuijen, C., Iriarte, M., Neyt, C., Sory, M. P., & Stainier, I. (1998). The virulence plasmid of Yersinia, an antihost genome. *Microbiol Mol Biol Rev, 62*(4), 1315-1352.

Csuros, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics, 26*(15), 1910-1912. doi:10.1093/bioinformatics/btq315

Cunneen, M. M., & Reeves, P. R. (2007). The Yersinia kristensenii O11 O-antigen gene cluster was acquired by lateral gene transfer and incorporated at a novel chromosomal locus. *Mol Biol Evol, 24*(6), 1355-1365. doi:10.1093/molbev/msm058

Darzentas, N. (2010). Circoletto: visualizing sequence similarity with Circos. *Bioinformatics, 26*(20), 2620-2621. doi:10.1093/bioinformatics/btq484

Daubin, V., Gouy, M., & Perriere, G. (2002). A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res, 12*(7), 1080-1090. doi:10.1101/gr.187002

de Almeida, A. M. P., Guiyoule, A., Guilvout, I., Iteman, I., Baranton, G., & Carniel, E. (1993). Chromosomal irp2 gene in Yersinia: distribution, expression, deletion and impact on virulence. *Microbial Pathogenesis, 14*(1), 9-21. doi:http://dx.doi.org/10.1006/mpat.1993.1002

de Queiroz, A., & Gatesy, J. (2007). The supermatrix approach to systematics. *Trends Ecol Evol, 22*(1), 34-41. doi:10.1016/j.tree.2006.10.002

Delcher, A. L., Phillippy, A., Carlton, J., & Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res, 30*(11), 2478-2483.

Desai, P. T., Porwollik, S., Long, F., Cheng, P., Wollam, A., Bhonagiri-Palsikar, V., Hallsworth-Pepin, K., Clifton, S. W., Weinstock, G. M., & McClelland, M. (2013). Evolutionary Genomics of Salmonella enterica Subspecies. *MBio, 4*(2). doi:10.1128/mBio.00579-12

Dewhirst, F. E., Shen, Z., Scimeca, M. S., Stokes, L. N., Boumenna, T., Chen, T., Paster, B. J., & Fox, J. G. (2005). Discordant 16S and 23S rRNA gene phylogenies for the genus Helicobacter: implications for phylogenetic inference and systematics. *J Bacteriol, 187*(17), 6106-6118. doi:10.1128/JB.187.17.6106-6118.2005

Didelot, X., & Wilson, D. J. (2015). ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol, 11*(2), e1004041. doi:10.1371/journal.pcbi.1004041

Eppinger, M., Rosovitz, M. J., Fricke, W. F., Rasko, D. A., Kokorina, G., Fayolle, C., Lindler, L. E., Carniel, E., & Ravel, J. (2007). The complete genome sequence of Yersinia pseudotuberculosis IP31758, the causative agent of Far East scarlet-like fever. *PLoS Genet, 3*(8), e142. doi:10.1371/journal.pgen.0030142

Espinosa-Urgel, M., & Kolter, R. (1998). Escherichia coli genes expressed preferentially in an aquatic environment. *Mol Microbiol, 28*(2), 325-332.

Ewing, W. H., Ross, A. J., Brenner, D. J., & Fanning, G. R. (1978). Yersinia ruckeri sp. nov., the Redmouth (RM) Bacterium. *International Journal of Systematic and Evolutionary Microbiology, 28*(1), 37-44. doi:doi:10.1099/00207713-28-1-37

Felek, S., Tsang, T. M., & Krukonis, E. S. (2010). Three Yersinia pestis adhesins facilitate Yop delivery to eukaryotic cells and contribute to plague virulence. *Infect Immun, 78*(10), 4134-4150. doi:10.1128/IAI.00167-10

Fernandez, L., Marquez, I., & Guijarro, J. A. (2004). Identification of specific in vivo-induced (ivi) genes in Yersinia ruckeri and analysis of ruckerbactin, a catecholate siderophore iron acquisition system. *Appl Environ Microbiol, 70*(9), 5199-5207. doi:10.1128/AEM.70.9.5199-5207.2004

Foultier, B., Troisfontaines, P., Muller, S., Opperdoes, F. R., & Cornelis, G. R. (2002). Characterization of the ysa pathogenicity locus in the chromosome of Yersinia enterocolitica and phylogeny analysis of type III secretion systems. *J Mol Evol, 55*(1), 37-51. doi:10.1007/s00239-001-0089-7

Fraser, C., Hanage, W. P., & Spratt, B. G. (2007). Recombination and the nature of bacterial speciation. *Science, 315*(5811), 476-480. doi:10.1126/science.1127573

Fredriksson-Ahomaa, M., Cernela, N., Hachler, H., & Stephan, R. (2012). Yersinia enterocolitica strains associated with human infections in Switzerland 2001-2010. *Eur J Clin Microbiol Infect Dis, 31*(7), 1543-1550. doi:10.1007/s10096-011-1476-7

Fukushima, H., Sato, T., Nagasako, R., & Takeda, I. (1991). Acute mesenteric lymphadenitis due to Yersinia pseudotuberculosis lacking a virulence plasmid. *J Clin Microbiol, 29*(6), 1271-1275.

Galindo, C. L., Rosenzweig, J. A., Kirtley, M. L., & Chopra, A. K. (2011). Pathogenesis of Y. enterocolitica and Y. pseudotuberculosis in Human Yersiniosis. *J Pathog, 2011*, 182051. doi:10.4061/2011/182051

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res, 43*(Database issue), D261-269. doi:10.1093/nar/gku1223

Garzetti, D., Bouabe, H., Heesemann, J., & Rakin, A. (2012). Tracing genomic variations in two highly virulent Yersinia enterocolitica strains with unequal ability to compete for host colonization. *BMC Genomics, 13*, 467. doi:10.1186/1471-2164-13-467

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res, 31*(13), 3784-3788.

Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., Liu, C., Shi, W., & Bryant, S. H. (2010). The NCBI BioSystems database. *Nucleic Acids Res, 38*(Database issue), D492-496. doi:10.1093/nar/gkp858

Georgiades, K., Merhej, V., El Karkouri, K., Raoult, D., & Pontarotti, P. (2011). Gene gain and loss events in Rickettsia and Orientia species. *Biol Direct, 6*, 6. doi:10.1186/1745-6150-6-6

Grant, T., Bennett-Wood, V., & Robins-Browne, R. M. (1998). Identification of virulence-associated characteristics in clinical isolates of Yersinia enterocolitica lacking classical virulence markers. *Infect Immun, 66*(3), 1113-1120.

Grassl, G. A., Bohn, E., Muller, Y., Buhler, O. T., & Autenrieth, I. B. (2003). Interaction of Yersinia enterocolitica with epithelial cells: invasin beyond invasion. *Int J Med Microbiol, 293*(1), 41-54. doi:10.1078/1438-4221-00243

Green, M. H., Jones, L., Little, L. K., Schamiloglu, U., & Sussman, G. D. (2014). Yersinia pestis and the three plague pandemics. *Lancet Infect Dis, 14*(10), 918. doi:10.1016/S1473-3099(14)70878-3

Haft, D. H., Selengut, J., Mongodin, E. F., & Nelson, K. E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol, 1*(6), e60. doi:10.1371/journal.pcbi.0010060

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., & Beck, E. (2013). TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res, 41*(Database issue), D387-395. doi:10.1093/nar/gks1234

Haller, J. C., Carlson, S., Pederson, K. J., & Pierson, D. E. (2000). A chromosomally encoded type III secretion pathway in Yersinia enterocolitica is important in virulence. *Mol Microbiol, 36*(6), 1436-1446.

Hammerl, J. A., Klein, I., Lanka, E., Appel, B., & Hertwig, S. (2008). Genetic and functional properties of the self-transmissible Yersinia enterocolitica plasmid pYE854, which mobilizes the virulence plasmid pYV. *J Bacteriol, 190*(3), 991-1010. doi:10.1128/JB.01467-07

Heesemann, J. (1987). Chromosomal-encoded siderophores are required for mouse virulence of enteropathogenic Yersinia species. *FEMS Microbiol Lett, 48*(1-2), 229-233. doi:10.1111/j.1574-6968.1987.tb02547.x

Hinnebusch, B. J., Rudolph, A. E., Cherepanov, P., Dixon, J. E., Schwan, T. G., & Forsberg, A. (2002). Role of Yersinia murine toxin in survival of Yersinia pestis in the midgut of the flea vector. *Science, 296*(5568), 733-735. doi:10.1126/science.1069972

Horvath, P., & Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. *Science, 327*(5962), 167-170. doi:10.1126/science.1179555

Howard, S. L., Gaunt, M. W., Hinds, J., Witney, A. A., Stabler, R., & Wren, B. W. (2006). Application of comparative phylogenomics to study the evolution of Yersinia enterocolitica and to identify genetic differences relating to pathogenicity. *J Bacteriol, 188*(10), 3645-3653. doi:10.1128/JB.188.10.3645-3653.2006

Hugouvieux-Cotte-Pattat, N., & Reverchon, S. (2001). Two transporters, TogT and TogMNAB, are responsible for oligogalacturonide uptake in Erwinia chrysanthemi 3937. *Mol Microbiol, 41*(5), 1125-1132.

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol, 23*(2), 254-267. doi:10.1093/molbev/msj030

Iriarte, M., & Cornelis, G. R. (1995). MyfF, an element of the network regulating the synthesis of fibrillae in Yersinia enterocolitica. *J Bacteriol, 177*(3), 738-744.

Iwobi, A., Heesemann, J., Garcia, E., Igwe, E., Noelting, C., & Rakin, A. (2003). Novel virulence-associated type II secretion system unique to high-pathogenicity Yersinia enterocolitica. *Infect Immun, 71*(4), 1872-1879.

Jensen, R. A. (2001). Orthologs and paralogs - we need to get it right. *Genome Biol, 2*(8), INTERACTIONS1002.

Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics, 11*, 431. doi:10.1186/1471-2105-11-431

Junier, T., & Zdobnov, E. M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics, 26*(13), 1669-1670. doi:10.1093/bioinformatics/btq243

Kanaujia, P. K., Bajaj, P., & Virdi, J. S. (2015). Analysis of iron acquisition and storage-related genes in clinical and non-clinical strains of Yersinia enterocolitica biovar 1A. *APMIS, 123*(10), 858-866. doi:10.1111/apm.12425

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol, 30*(4), 772-780. doi:10.1093/molbev/mst010

Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G. M., Richardson, P., & Chisholm, S. W. (2007). Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. *PLoS Genet, 3*(12), e231. doi:10.1371/journal.pgen.0030231

Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., & Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biol, 3*(2), RESEARCH0008.

Konstantinidis, K. T., & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A, 102*(7), 2567-2572. doi:10.1073/pnas.0409727102

Kopac, S., Wang, Z., Wiedenbeck, J., Sherry, J., Wu, M., & Cohan, F. M. (2014). Genomic heterogeneity and ecological speciation within one subspecies of Bacillus subtilis. *Appl Environ Microbiol, 80*(16), 4842-4853. doi:10.1128/AEM.00576-14

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res, 19*(9), 1639-1645. doi:10.1101/gr.092759.109

Lapierre, P., Lasek-Nesselquist, E., & Gogarten, J. P. (2014). The impact of HGT on phylogenomic reconstruction methods. *Brief Bioinform, 15*(1), 79-90. doi:10.1093/bib/bbs050

Lassalle, F., Muller, D., & Nesme, X. (2015). Ecological speciation in bacteria: reverse ecology approaches reveal the adaptive part of bacterial cladogenesis. *Res Microbiol, 166*(10), 729-741. doi:10.1016/j.resmic.2015.06.008

Lathem, W. W., Price, P. A., Miller, V. L., & Goldman, W. E. (2007). A plasminogen-activating protease specifically controls the development of primary pneumonic plague. *Science, 315*(5811), 509-513. doi:10.1126/science.1137195

Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics, 12*, 124. doi:10.1186/1471-2105-12-124

Lee, E. J., Cho, Y. H., Kim, H. S., Ahn, B. E., & Roe, J. H. (2004). Regulation of sigmaB by an anti- and an anti-anti-sigma factor in Streptomyces coelicolor in response to osmotic stress. *J Bacteriol, 186*(24), 8490-8498. doi:10.1128/JB.186.24.8490-8498.2004

Leong, J. M., Fournier, R. S., & Isberg, R. R. (1990). Identification of the integrin binding domain of the Yersinia pseudotuberculosis invasin protein. *EMBO J, 9*(6), 1979-1989.

Lerat, E., & Ochman, H. (2004). Psi-Phi: exploring the outer limits of bacterial pseudogenes. *Genome Res, 14*(11), 2273-2278. doi:10.1101/gr.2925604

Letoffe, S., Nato, F., Goldberg, M. E., & Wandersman, C. (1999). Interactions of HasA, a bacterial haemophore, with haemoglobin and with its outer membrane receptor HasR. *Mol Microbiol, 33*(3), 546-555.

Lian, C. J., Hwang, W. S., Kelly, J. K., & Pai, C. H. (1987). Invasiveness of Yersinia enterocolitica lacking the virulence plasmid: an in-vivo study. *J Med Microbiol, 24*(3), 219-226. doi:10.1099/00222615-24-3-219

Maier, R. J., Olczak, A., Maier, S., Soni, S., & Gunn, J. (2004). Respiratory hydrogen use by Salmonella enterica serovar Typhimurium is essential for virulence. *Infect Immun, 72*(11), 6294-6299. doi:10.1128/IAI.72.11.6294-6299.2004

Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M., & Dowson, C. G. (2000). Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. *J Bacteriol, 182*(4), 1016-1023.

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J., Wolf, Y. I., Yakunin, A. F., van der Oost, J., & Koonin, E. V. (2011). Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol, 9*(6), 467-477. doi:10.1038/nrmicro2577

Markowitz, V. M., Chen, I. M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., Huntemann, M., Anderson, I., Mavromatis, K., Ivanova, N. N., & Kyrpides, N. C. (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res, 40*(Database issue), D115-122. doi:10.1093/nar/gkr1044

Marraffini, L. A. (2013). CRISPR-Cas immunity against phages: its effects on the evolution and survival of bacterial pathogens. *PLoS Pathog, 9*(12), e1003765. doi:10.1371/journal.ppat.1003765

Marraffini, L. A., & Sontheimer, E. J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science, 322*(5909), 1843-1845. doi:10.1126/science.1165771

Marri, P. R., Hao, W., & Golding, G. B. (2007). The role of laterally transferred genes in adaptive evolution. *BMC Evol Biol, 7 Suppl 1*, S8. doi:10.1186/1471-2148-7-S1-S8

McDonald, C., Vacratsis, P. O., Bliska, J. B., & Dixon, J. E. (2003). The yersinia virulence factor YopM forms a novel protein complex with two cellular kinases. *J Biol Chem, 278*(20), 18514-18523. doi:10.1074/jbc.M301226200

McNally, A., Thomson, N. R., Reuter, S., & Wren, B. W. (2016). 'Add, stir and reduce': Yersinia spp. as model bacteria for pathogen evolution. *Nat Rev Microbiol, 14*(3), 177-190. doi:10.1038/nrmicro.2015.29

Merhej, V., Adekambi, T., Pagnier, I., Raoult, D., & Drancourt, M. (2008a). Yersinia massiliensis sp. nov., isolated from fresh water. *Int J Syst Evol Microbiol, 58*(Pt 4), 779-784. doi:10.1099/ijs.0.65219-0

Merhej, V., Adékambi, T., Pagnier, I., Raoult, D., & Drancourt, M. (2008b). Yersinia massiliensis sp. nov., isolated from fresh water. *International Journal of Systematic and Evolutionary Microbiology, 58*(4), 779-784. doi:doi:10.1099/ijs.0.65219-0

Mikula, K. M., Kolodziejczyk, R., & Goldman, A. (2012). Yersinia infection tools-characterization of structure and function of adhesins. *Front Cell Infect Microbiol, 2*, 169. doi:10.3389/fcimb.2012.00169

Miller, V. L., Farmer, J. J., 3rd, Hill, W. E., & Falkow, S. (1989). The ail locus is found uniquely in Yersinia enterocolitica serotypes commonly associated with disease. *Infect Immun, 57*(1), 121-131.

Mirold, S., Rabsch, W., Rohde, M., Stender, S., Tschape, H., Russmann, H., Igwe, E., & Hardt, W. D. (1999). Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic Salmonella typhimurium strain. *Proc Natl Acad Sci U S A, 96*(17), 9845-9850.

Mobley, H. L. (1996). The role of Helicobacter pylori urease in the pathogenesis of gastritis and peptic ulceration. *Aliment Pharmacol Ther, 10 Suppl 1*, 57-64.

Navarro, L., Alto, N. M., & Dixon, J. E. (2005). Functions of the Yersinia effector proteins in inhibiting host immune responses. *Curr Opin Microbiol, 8*(1), 21-27. doi:10.1016/j.mib.2004.12.014

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., McDonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., White, O., Salzberg, S. L., Smith, H. O., Venter, J. C., & Fraser, C. M. (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature, 399*(6734), 323-329. doi:10.1038/20601

Neubauer, H., Aleksic, S., Hensel, A., Finke, E. J., & Meyer, H. (2000). Yersinia enterocolitica 16S rRNA gene types belong to the same genospecies but form three homology groups. *Int J Med Microbiol, 290*(1), 61-64. doi:10.1016/S1438-4221(00)80107-1

Neyt, C., Iriarte, M., Thi, V. H., & Cornelis, G. R. (1997). Virulence and arsenic resistance in Yersiniae. *J Bacteriol, 179*(3), 612-619.

Nishi, J., Sheikh, J., Mizuguchi, K., Luisi, B., Burland, V., Boutin, A., Rose, D. J., Blattner, F. R., & Nataro, J. P. (2003). The export of coat protein from enteroaggregative Escherichia coli by a specific ATP-binding cassette transporter system. *J Biol Chem, 278*(46), 45680-45689. doi:10.1074/jbc.M306413200

Nozawa, T., Furukawa, N., Aikawa, C., Watanabe, T., Haobam, B., Kurokawa, K., Maruyama, F., & Nakagawa, I. (2011). CRISPR inhibition of prophage acquisition in Streptococcus pyogenes. *PLoS One, 6*(5), e19543. doi:10.1371/journal.pone.0019543

Ochman, H., & Davalos, L. M. (2006). The nature and dynamics of bacterial genomes. *Science, 311*(5768), 1730-1733. doi:10.1126/science.1119966

Ochman, H., Lawrence, J. G., & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature, 405*(6784), 299-304. doi:10.1038/35012500

Pal, C., Papp, B., & Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet, 37*(12), 1372-1375. doi:10.1038/ng1686

Parkhill, J., Wren, B. W., Thomson, N. R., Titball, R. W., Holden, M. T., Prentice, M. B., Sebaihia, M., James, K. D., Churcher, C., Mungall, K. L., Baker, S., Basham, D., Bentley, S. D., Brooks, K., Cerdeno-Tarraga, A. M., Chillingworth, T., Cronin, A., Davies, R. M., Davis, P., Dougan, G., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Karlyshev, A. V., Leather, S., Moule, S., Oyston, P. C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S., & Barrell, B. G. (2001). Genome sequence of Yersinia pestis, the causative agent of plague. *Nature, 413*(6855), 523-527. doi:10.1038/35097083

Pelludat, C., Hogardt, M., & Heesemann, J. (2002). Transfer of the Core Region Genes of the Yersinia enterocolitica WA-C Serotype O:8 High-Pathogenicity Island to Y. enterocolitica MRS40, a Strain with Low Levels of Pathogenicity, Confers a Yersiniabactin Biosynthesis Phenotype and Enhanced Mouse Virulence. *Infect Immun, 70*(4), 1832-1841. doi:10.1128/iai.70.4.1832-1841.2002

Pelludat, C., Rakin, A., Jacobi, C. A., Schubert, S., & Heesemann, J. (1998). The yersiniabactin biosynthetic gene cluster of Yersinia enterocolitica: organization and siderophore-dependent regulation. *J Bacteriol, 180*(3), 538-546.

Perry, R. D., & Fetherston, J. D. (1997). Yersinia pestis--etiologic agent of plague. *Clin Microbiol Rev, 10*(1), 35-66.

Peterson, J. W. (1996). Bacterial Pathogenesis. In S. Baron (Ed.), *Medical Microbiology* (4th ed.). Galveston (TX).

Ponnusamy, D., & Clinkenbeard, K. D. (2015). Role of Tellurite Resistance Operon in Filamentous Growth of Yersinia pestis in Macrophages. *PLoS One, 10*(11), e0141984. doi:10.1371/journal.pone.0141984

Ponnusamy, D., Hartson, S. D., & Clinkenbeard, K. D. (2011). Intracellular Yersinia pestis expresses general stress response and tellurite resistance proteins in mouse macrophages. *Vet Microbiol, 150*(1-2), 146-151. doi:10.1016/j.vetmic.2010.12.025

Pontes, M. H., Lee, E. J., Choi, J., & Groisman, E. A. (2015). Salmonella promotes virulence by repressing cellulose production. *Proc Natl Acad Sci U S A, 112*(16), 5183-5188. doi:10.1073/pnas.1500989112

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One, 5*(3), e9490. doi:10.1371/journal.pone.0009490

Price-Carter, M., Tingey, J., Bobik, T. A., & Roth, J. R. (2001). The alternative electron acceptor tetrathionate supports B12-dependent anaerobic growth of Salmonella enterica serovar typhimurium on ethanolamine or 1,2-propanediol. *J Bacteriol, 183*(8), 2463-2475. doi:10.1128/JB.183.8.2463-2475.2001

Rajendhran, J., & Gunasekaran, P. (2011). Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. *Microbiol Res, 166*(2), 99-110. doi:10.1016/j.micres.2010.02.003

Rakin, A., Schneider, L., & Podladchikova, O. (2012). Hunger for iron: the alternative siderophore iron scavenging systems in highly virulent Yersinia. *Front Cell Infect Microbiol, 2*, 151. doi:10.3389/fcimb.2012.00151

Ravenhall, M., Skunca, N., Lassalle, F., & Dessimoz, C. (2015). Inferring horizontal gene transfer. *PLoS Comput Biol, 11*(5), e1004095. doi:10.1371/journal.pcbi.1004095

Reuter, S., Connor, T. R., Barquist, L., Walker, D., Feltwell, T., Harris, S. R., Fookes, M., Hall, M. E., Petty, N. K., Fuchs, T. M., Corander, J., Dufour, M., Ringwood, T., Savin, C., Bouchier, C., Martin, L., Miettinen, M., Shubin, M., Riehm, J. M., Laukkanen-Ninios, R., Sihvonen, L. M., Siitonen, A., Skurnik, M., Falcao, J. P., Fukushima, H., Scholz, H. C., Prentice, M. B., Wren, B. W., Parkhill, J., Carniel, E., Achtman, M., McNally, A., & Thomson, N. R. (2014). Parallel independent evolution of pathogenicity within the genus Yersinia. *Proc Natl Acad Sci U S A, 111*(18), 6768-6773. doi:10.1073/pnas.1317161111

Richter, M., & Rossello-Mora, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A, 106*(45), 19126-19131. doi:10.1073/pnas.0906412106

Rohmer, L., Hocquet, D., & Miller, S. I. (2011). Are pathogenic bacteria just looking for food? Metabolism and microbial pathogenesis. *Trends Microbiol, 19*(7), 341-348. doi:10.1016/j.tim.2011.04.003

Romalde, J. L., & Toranzo, A. E. (1993). Pathological activities of Yersinia ruckeri, the enteric redmouth (ERM) bacterium. *FEMS Microbiol Lett, 112*(3), 291-299.

Ross, A. J., Rucker, R. R., & Ewing, W. H. (1966). Description of a bacterium associated with redmouth disease of rainbow trout (Salmo gairdneri). *Can J Microbiol, 12*(4), 763-770.

Roy, C., Kester, H., Visser, J., Shevchik, V., Hugouvieux-Cotte-Pattat, N., Robert-Baudouy, J., & Benen, J. (1999). Modes of action of five different endopectate lyases from Erwinia chrysanthemi 3937. *J Bacteriol, 181*(12), 3705-3709.

San Millan, A., Pena-Miller, R., Toll-Riera, M., Halbert, Z. V., McLean, A. R., Cooper, B. S., & MacLean, R. C. (2014). Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. *Nat Commun, 5*, 5208. doi:10.1038/ncomms6208

Sasikaran, J., Ziemski, M., Zadora, P. K., Fleig, A., & Berg, I. A. (2014). Bacterial itaconate degradation promotes pathogenicity. *Nat Chem Biol, 10*(5), 371-377. doi:10.1038/nchembio.1482

Schaake, J., Drees, A., Gruning, P., Uliczka, F., Pisano, F., Thiermann, T., von Altrock, A., Seehusen, F., Valentin-Weigand, P., & Dersch, P. (2014). Essential role of invasin for colonization and persistence of Yersinia enterocolitica in its natural reservoir host, the pig. *Infect Immun, 82*(3), 960-969. doi:10.1128/IAI.01001-13

Segata, N., & Huttenhower, C. (2011). Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. *PLoS One, 6*(9), e24704. doi:10.1371/journal.pone.0024704

Segerman, B. (2012). The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. *Front Cell Infect Microbiol, 2*, 116. doi:10.3389/fcimb.2012.00116

Simonet, M., Riot, B., Fortineau, N., & Berche, P. (1996). Invasin production by Yersinia pestis is abolished by insertion of an IS200-like element within the inv gene. *Infect Immun, 64*(1), 375-379.

Singhal, N., Kumar, M., & Virdi, J. S. (2016). Resistance to amoxicillin-clavulanate and its relation to virulence-related factors in Yersinia enterocolitica biovar 1A. *Indian J Med Microbiol, 34*(1), 85-87. doi:10.4103/0255-0857.174125

Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J., & Holmes, I. H. (2009). JBrowse: a next-generation genome browser. *Genome Res, 19*(9), 1630-1638. doi:10.1101/gr.094607.109

Skurnik, M., & Wolf-Watz, H. (1989). Analysis of the yopA gene encoding the Yop1 virulence determinants of Yersinia spp. *Mol Microbiol, 3*(4), 517-529.

Snel, B., Bork, P., & Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nat Genet, 21*(1), 108-110. doi:10.1038/5052

Snel, B., Bork, P., & Huynen, M. A. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res, 12*(1), 17-25. doi:10.1101/gr.176501

Sonnhammer, E. L., & Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet, 18*(12), 619-620.

Sprague, L. D., & Neubauer, H. (2014). Genome Sequence of Yersinia similis Y228T, a Member of the Yersinia pseudotuberculosis Complex. *Genome Announc, 2*(2). doi:10.1128/genomeA.00216-14

Stahl, M., Friis, L. M., Nothaft, H., Liu, X., Li, J., Szymanski, C. M., & Stintzi, A. (2011). L-fucose utilization provides Campylobacter jejuni with a competitive advantage. *Proc Natl Acad Sci U S A, 108*(17), 7194-7199. doi:10.1073/pnas.1014125108

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313. doi:10.1093/bioinformatics/btu033

Stenkova, A. M., Isaeva, M. P., Bystritskaya, E. P., Guzev, K. V., Rasskazov, V. A., & Rakin, A. (2012). The molecular phylogeny of the gyrB gene: a molecular marker for systematic characterization of the genus Yersinia. *Adv Exp Med Biol, 954*, 53-56. doi:10.1007/978-1-4614-3561-7_7

Stephan, R., Joutsen, S., Hofer, E., Sade, E., Bjorkroth, J., Ziegler, D., & Fredriksson-Ahomaa, M. (2013). Characteristics of Yersinia enterocolitica biotype 1A strains isolated from patients and asymptomatic carriers. *Eur J Clin Microbiol Infect Dis, 32*(7), 869-875. doi:10.1007/s10096-013-1820-1

Sulakvelidze, A. (2000). Yersiniae other than Y. enterocolitica, Y. pseudotuberculosis, and Y. pestis: the ignored species. *Microbes Infect, 2*(5), 497-513.

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res, 34*(Web Server issue), W609-612. doi:10.1093/nar/gkl315

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol, 30*(12), 2725-2729. doi:10.1093/molbev/mst197

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., Deboy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., Brinkac, L. M., Dodson, R. J., Rosovitz, M. J., Sullivan, S. A., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, M. L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K. J., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Kasper, D. L., Telford, J. L., Wessels, M. R., Rappuoli, R., & Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A, 102*(39), 13950-13955. doi:10.1073/pnas.0506758102

Thomson, N. R., Howard, S., Wren, B. W., Holden, M. T., Crossman, L., Challis, G. L., Churcher, C., Mungall, K., Brooks, K., Chillingworth, T., Feltwell, T., Abdellah, Z., Hauser, H., Jagels, K., Maddison, M., Moule, S., Sanders, M., Whitehead, S., Quail, M. A., Dougan, G., Parkhill, J., & Prentice, M. B. (2006). The complete genome sequence and comparative genome analysis of the high pathogenicity Yersinia enterocolitica strain 8081. *PLoS Genet, 2*(12), e206. doi:10.1371/journal.pgen.0020206

Tonini, J., Moore, A., Stern, D., Shcheglovitova, M., & Orti, G. (2015). Concatenation and Species Tree Methods Exhibit Statistically Indistinguishable Accuracy under a Range of Simulated Conditions. *PLoS Curr, 7*. doi:10.1371/currents.tol.34260cc27551a527b124ec5f6334b6be

Uchiyama, I., Mihara, M., Nishide, H., & Chiba, H. (2013). MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. *Nucleic Acids Res, 41*(Database issue), D631-635. doi:10.1093/nar/gks1006

Valentin-Weigand, P., Heesemann, J., & Dersch, P. (2014). Unique virulence properties of Yersinia enterocolitica O:3--an emerging zoonotic pathogen using pigs as preferred reservoir host. *Int J Med Microbiol, 304*(7), 824-834. doi:10.1016/j.ijmm.2014.07.008

von Haeseler, A. (2012). Do we still need supertrees? *BMC Biol, 10*, 13. doi:10.1186/1741-7007-10-13

Wagner, A. (2002). Selection and gene duplication: a view from the genome. *Genome Biol, 3*(5), reviews1012.

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E., Vonstein, V., Warren, A., Will, R., Wilson, M. J., Yoo, H. S., Zhang, C., Zhang, Y., & Sobral, B. W. (2014). PATRIC, the bacterial bioinformatics

database and analysis resource. *Nucleic Acids Res, 42*(Database issue), D581-591. doi:10.1093/nar/gkt1099

Wauters, G., Kandolo, K., & Janssens, M. (1987). Revised biogrouping scheme of Yersinia enterocolitica. *Contrib Microbiol Immunol, 9*, 14-21.

Wiedenbeck, J., & Cohan, F. M. (2011). Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev, 35*(5), 957-976. doi:10.1111/j.1574-6976.2011.00292.x

Williams, K. P., Gillespie, J. J., Sobral, B. W., Nordberg, E. K., Snyder, E. E., Shallom, J. M., & Dickerman, A. W. (2010). Phylogeny of gammaproteobacteria. *J Bacteriol, 192*(9), 2305-2314. doi:10.1128/JB.01480-09

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A, 87*(12), 4576-4579.

Wren, B. W. (2003). The yersiniae--a model genus to study the rapid evolution of bacterial pathogens. *Nat Rev Microbiol, 1*(1), 55-64. doi:10.1038/nrmicro730

Wu, M. C., Chen, Y. C., Lin, T. L., Hsieh, P. F., & Wang, J. T. (2012). Cellobiose-specific phosphotransferase system of Klebsiella pneumoniae and its importance in biofilm formation and virulence. *Infect Immun, 80*(7), 2464-2472. doi:10.1128/IAI.06247-11

Xavier, K. B., Miller, S. T., Lu, W., Kim, J. H., Rabinowitz, J., Pelczer, I., Semmelhack, M. F., & Bassler, B. L. (2007). Phosphorylation and processing of the quorum-sensing molecule autoinducer-2 in enteric bacteria. *ACS Chem Biol, 2*(2), 128-136. doi:10.1021/cb600444h

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C. Y., & Wei, L. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res, 39*(Web Server issue), W316-322. doi:10.1093/nar/gkr483

Xu, Y., Zhu, Y., Wang, Y., Chang, Y. F., Zhang, Y., Jiang, X., Zhuang, X., Zhu, Y., Zhang, J., Zeng, L., Yang, M., Li, S., Wang, S., Ye, Q., Xin, X., Zhao, G., Zheng, H., Guo, X., & Wang, J. (2016). Whole genome sequencing revealed host adaptation-focused genomic plasticity of pathogenic Leptospira. *Sci Rep, 6*, 20020. doi:10.1038/srep20020

Yamazaki, A., Li, J., Hutchins, W. C., Wang, L., Ma, J., Ibekwe, A. M., & Yang, C. H. (2011). Commensal effect of pectate lyases secreted from Dickeya dadantii on proliferation of Escherichia coli O157:H7 EDL933 on lettuce leaves. *Appl Environ Microbiol, 77*(1), 156-162. doi:10.1128/AEM.01079-10

Yang, J., Chen, L., Sun, L., Yu, J., & Jin, Q. (2008). VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res, 36*(Database issue), D539-542. doi:10.1093/nar/gkm951

Yang, Y., Merriam, J. J., Mueller, J. P., & Isberg, R. R. (1996). The psa locus is responsible for thermoinducible binding of Yersinia pseudotuberculosis to cultured cells. *Infect Immun, 64*(7), 2483-2489.

Yaron, S., & Romling, U. (2014). Biofilm formation by enteric pathogens and its role in plant colonization and persistence. *Microb Biotechnol, 7*(6), 496-516. doi:10.1111/1751-7915.12186

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., & Brinkman, F. S. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics, 26*(13), 1608-1615. doi:10.1093/bioinformatics/btq249

LIST OF PUBLICATIONS AND PAPERS PRESENTED

Tan, S. Y., Dutta, A., Jakubovics, N. S., Ang, M. Y., Siow, C. C., Mutha, N. V., Heydari, H., Wee, W. Y., Wong, G. J., & Choo, S. W. (2015). YersiniaBase: a genomic resource and analysis platform for comparative analysis of Yersinia. BMC Bioinformatics, 16, 9. doi: 10.1186/s12859-014-0422-y

Tan, S. Y., Tan, I. K. P., Tan, M. F., Dutta, A., & Choo, S. W. (2016). Evolutionary study of Yersinia genomes deciphers emergence of human pathogenic species. Sci. Rep. 6, 36116. doi: 10.1038/srep36116